UNIVERSIDAD CARLOS III

ESCUELA POLITÉCNICA SUPERIOR

GRADO EN INGENIERÍA INFORMÁTICA

Análisis de Información Proveniente de Redes Sociales como Twitter

---

Autor: Soledad Martín Morales
Tutora: Ana María Iglesias Maqueda

Junio 2014

RESUMEN

Nos encontramos ante un momento en el que las redes sociales se han hecho imprescindibles en el uso diario de nuestras vidas.

Actualmente existe una diversidad de plataformas gratuitas en la web en las que puedes registrarte y compartir estados, fotos, pensamientos, etc. Gracias a ello, estas plataformas recogen millones de datos de todas las personas, y una vez procesados reportan información muy útil, entre otras muchas cosas, para estudios de mercado.

El problema con el que nos encontramos es que todos estos datos no siempre vienen estructurados y sería necesario que siguieran un patrón para su uso eficiente. Para obtener esta información tanto de la Web como de las Redes Sociales es necesario realizar un proceso de extracción, normalización y análisis exhaustivo de los mismos.

En este proyecto he realizado el análisis del sentimiento (positivo, negativo o neutro) de un conjunto de textos obtenidos de Twitter mediante los procesos expuestos anteriormente.

Palabras Clave

Redes Sociales, Twitter, Tweets, Análisis de Sentimiento, Minería de Datos (Data Mining)

ABSTRACT

We are facing a time when social media have become indispensable in daily use in our lives.

Currently, a great variety of free web platforms and social networks are available where everybody can register and share statements, pictures, thoughts, etc.. As a result, these platforms collect millions of data of all people, and once processed, it could report very useful information, for example, for marketing issues and business intelligence.

The problem that we face is that these data are not always structured in order to be efficient in their use. Then, some steps are needed in order to be able to process information from the Web and Social Networks: data extraction, normalization and thorough analysis of these data.

This project is focused in detecting the sentiments (positive, negative or neutral) of a set Twittertweets. A process of extraction, normalization and analysis is carried.

Key Word

Twitter, Tweets, Sentiment Analysis, Data Mining

AGRADECIMIENTOS

Llegado este momento en el que finaliza una de las etapas más importantes de mi vida, son muchos los nombres que se me pasan por la cabeza.

En primer lugar quiero dar las gracias a mi familia, y en especial a mi madre que día tras día se asomaba a mi habitación para “ver lo bien que llevaba el examen” o “ver todo lo que había avanzado con el proyecto”. Todos, en todo momento, han estado mostrándome el apoyo que se necesita para no tirar la toalla en los momentos de estrés.

Por supuesto a Berni, compañero durante todos los años de carrera y desde hace dos años y ya para toda la vida, de una forma más especial. Con él he sufrido el estrés en todas y cada una de las asignaturas, esos nervios pre-exámenes, los días de biblioteca, las prácticas interminables…

A mis amigas, por ser ese apoyo incondicional en cualquier momento y siempre saben sacarte una sonrisa.

Quiero también agradecer a mis compañeros del trabajo, que en este último año me han dado toda clase de facilidades para poder acabar la carrera y poder entrar el proyecto. Además, cada día se interesaban por cómo iba y me ofrecían su ayuda.

También quiero agradecer a todos y cada uno de los profesores que he tenido a lo largo de la carrera, porque de todos he aprendido algo.

Y he querido dejar para el final a la persona que me ha ayudado durante estos últimos meses de Universidad. Ana, sin ti entregar el proyecto este año no habría sido posible. Gracias por esa predisposición continua, todos tus consejos y tus ganas de trabajar y de que todo salga perfecto.

Gracias a todas y cada una de las personas que han formado parte de mi vida durante todos estos años.

CONTENIDO

TABLA DE ILUSTRACIONES

1 INTRODUCCIÓN

A lo largo de este documento se va a realizar un análisis del estado actual de la cuestión, analizando diferentes sistemas capaces de extraer información de las diferentes redes sociales. Así mismo, se estudiará la forma de analizar los datos extraídos para convertirlos en información útil.

Habrá una descripción de los principales métodos de obtención y tratamiento de esa información y, en concreto, la explicación de cómo se ha creado un motor que recoge información a partir de Twitter y la transforma de tal manera que se pueda obtener un análisis de sentimientos (positivo/negativo) de estos.

Cuando se habla de sentimiento de un tweet se hace referencia a la clasificación de un ente (objeto, persona, marca, etc.) a través de la valoración positiva, negativa o neutra del contenido del texto expuesto en Twitter. Al realizar una clasificación de todos los tweets relacionados con este ente, se podrá obtener un balance general basado en cifras exactas de lo que los usuarios de Twitter piensan sobre él. Se utiliza sobre todo para estrategia de mercado.

1.1 CONTEXTO

Para contextualizar mi proyecto veo necesario hablar de la Web 2.0, la Web Semántica y los inicios de la Web 3.0 con la influencia que las anteriores están teniendo en la actualidad.

Si nos remontamos al año 2000 la web era únicamente una herramienta de trabajo, sobre todo para científicos, y ahora forma parte de la vida de más de mil millones de usuarios. Además, cada vez se hace más indispensable buscar información de un producto, buscar cómo llegar a un lugar, elegir un destino de vacaciones… para todas estas tareas rutinarias hacemos uso de esta tecnología.

Podemos entender la Web 2.0 como una actitud más que como una tecnología, se crea una tendencia de aplicaciones web centradas en el usuario final. Estos usuarios dejan de ser usuarios pasivos (solo recibir información) a ser usuarios activos (tener capacidad de generar y compartir información con el resto de usuarios). Se crean portales de blogging y se mejora la accesibilidad.

La Web Semántica propone introducir descripciones explícitas sobre el significado de los recursos, para permitir que las propias máquinas tengan un nivel de comprensión de la web suficiente como para hacerse cargo de una parte, la más costosa, rutinaria, o físicamente inabarcable, del trabajo que actualmente realizan manualmente los usuarios que navegan e interactúan con la web. Es una corriente promovida por el propio inventor de la web y presidente del consorcio W3C cuyo último fin es lograr que las máquinas puedan entender, y por tanto utilizar, lo que la web contiene. Esta nueva web estaría poblada por agentes o representantes software capaces de navegar y realizar operaciones por nosotros para ahorrarnos trabajo y optimizar los resultados. Para conseguir esta meta, la web semántica propone describir los recursos de la web con representaciones procesables (es decir,

entendibles) no sólo por personas, sino por programas que puedan asistir, representar, o reemplazar a las personas en tareas rutinarias o inabarcables para un humano. Las tecnologías de la web semántica buscan desarrollar una web más cohesionada, donde sea aún más fácil localizar, compartir e integrar información y servicios, para sacar un partido todavía mayor de los recursos disponibles en la web (Castells, 2012) (Pascual, 2012).

Se podría hablar del nacimiento de la Web 3.0 por la convergencia de la Web Semántica y la Web 2.0. La diferencia de la Web 3.0 con la anterior es que además de poder opinar, se realizará el tratamiento de esa información y se podrá personalizar el contenido al usuario que lo esté solicitando. Se añadirá contenido semántico a los documentos y las máquinas que analicen los datos se basarán en los perfiles de los usuarios para responder a las solicitudes. La gestión de la información se hace en la nube y puede ser ejecutada desde cualquier dispositivo.

Como parte de las acciones realizadas en la Web 3.0 se incluye este proyecto. Ya que un usuario plasma su opinión en Twitter, éste recoge la información, la distribuye a partir de su API, el motor la analiza y se obtiene el conocimiento del sentimiento a partir de ella.

1.2 MOTIVACIÓN

El principal motivo que me ha llevado a realizar un trabajo relacionado con el mundo de las redes sociales y el análisis de contenido es conocer y aportar mi granito de arena en la gestión de la gran cantidad de datos que podemos encontrar en la red.

Me siento muy familiarizada con todo el tema de la Web 3.0, el compartir información en diferentes páginas, realizar compras por internet, y me pareció muy interesante elegir un proyecto en el que pudiese aprender cómo influyen todas las acciones que realiza un usuario, en la toma de decisiones de otra persona interesada en esa información.

Uno de los portales más usados a día de hoy por las empresas para adquirir toda esa información es Twitter por lo que ha sido la base de la información de mi proyecto.

1.3 OBJETIVOS DEL PROYECTO

Se puede dividir el proyecto en dos partes: la obtención de tweets mediante el API de Twitter y el análisis de los tweets obtenidos.

1.3.1 OBJETIVO PRINCIPAL

El objetivo principal de este proyecto es la creación de un motor que analice los tweets obtenidos a través de Twitter valorando si el sentimiento es positivo, negativo o neutro. Este análisis es realizado con una serie de herramientas obtenidas de manera externa o han sido creadas de forma autónoma.

Como se ha comentado anteriormente se puede dividir en dos fases definidas de la siguiente manera:

- Obtención de tweets a partir del API Streaming de Twitter.
- Análisis de tweets. Además este análisis está dividido en:
  - ✓ Extracción de tokens.
  - ✓ Extracción de Smileys y Hashtags.
  - ✓ Análisis y normalización de los tuits mediante un analizador de textos desarrollado la RAE.
  - ✓ Análisis del texto mediante los lematizadores de Apicultur.
  - ✓ Establecimiento de valores de cada token.
  - ✓ Generación de sentimiento.

1.3.2 OBJETIVOS SECUNDARIOS

Como objetivos secundarios se podrían considerar la generación de estadísticas y la elaboración de informes con los datos obtenidos del motor.

1.4 ESTRUCTURA DEL DOCUMENTO

En primer lugar se ofrece un pequeño resumen del proyecto tanto en español como en inglés y se dedican unas palabras de agradecimiento a toda la gente que ha hecho posible que esto salga adelante.

A partir del índice del documento, la estructura es la siguiente:

1. Introducción: en ella se realiza una breve prólogo del proyecto, la motivación que me ha llevado a realizarlo, los objetivos y la estructura del documento.
2. Estado del arte: es uno de los apartados más extensos del documento y donde se han invertido muchas horas de trabajo. Se realiza un estudio en profundidad tanto de las redes sociales como del estado del arte de la minería de datos y de las herramientas utilizadas para analizar el sentimiento del contenido de la red.
3. Ciclo de vida del proyecto: se describe el ciclo de vida que ha llevado el proyecto.
4. Metodología: se define la metodología utilizada en el proyecto.
5. Análisis del sistema: Se analizan los requisitos de usuario y las herramientas utilizadas en la aplicación.
6. Diseño e Implementación del Sistema: Se presenta la arquitectura de la aplicación y se detallan cada una de las fases de éste.
7. Pruebas: Se muestran las pruebas realizadas por cada una de las herramientas que componen el motor (pruebas unitarias) y las pruebas generales del motor.
8. Planificación: Se define la planificación estimada y la planificación real, en ambas se ha diseñado un diagrama de horas y un diagrama de Gant con las fechas.
9. Presupuesto: Al igual que en la planificación se define un presupuesto inicial, uno final y las diferencias que han existido entre uno y otro.
10. Protección de datos: Se indica el cumplimiento de la ley.
11. Conclusiones: Se detallan las conclusiones del trabajo y un breve resumen de éste.
12. Trabajo Futuro: Se describen las mejoras que se pueden realizar en el proyecto que por falta de tiempo o de herramientas no ha sido posible implantarlas en esta etapa.
13. Diccionario de Acrónimos: Diccionario con los acrónimos utilizados en el documento.

14. Bibliografía: Se realiza un índice con todos los documentos y los textos que se han utilizado en el documento.
15. ANEXO I: Cálculo Salario Profesionales: se muestran los cálculos realizados para definir el salario de cada profesional.

2 ESTADO DEL ARTE

En este apartado se describe de forma teórica la base en la que se sustenta este Trabajo. En primer lugar se hará una pequeña introducción con un análisis del uso de las redes sociales en la actualidad, a continuación se describirá brevemente en qué consiste la extracción de datos (Data Mining) y cómo es el proceso de extracción y procesamiento de datos, y por último se describirán algunas de las herramientas que se pueden utilizar tanto para clasificar el texto de modo que se pueda estudiar cómo para realizar un análisis con esa información obtenida.

2.1 FUENTES DE DATOS

El núcleo de la obtención de los datos se encuentra en las redes sociales. Se entiende como red social un sitio web que permite a los usuarios mantener comunicaciones con otros usuarios, expresar sus opiniones, generación de grupos con intereses similares, etc. Hay redes sociales de diversos ámbitos: redes personales, redes laborales, redes temáticas…

A continuación se van a detallar tres de ellas que en mi opinión tienen más relevancia en la actualidad: Twitter, Facebook y LinkedIn.

2.1.1 Twitter

Twitter es un servicio de Microblogging creado por Jack Dorsey en 2006, aunque ahora está bajo la jurisdicción de Daleware.

Se estima que en la actualidad tiene más de 500 millones de usuarios y genera alrededor de 65 millones de tweets diarios. Cabe destacar que esta aplicación es usada por grandes figuras públicas como el presidente de los Estados Unidos Barack Obama y actores y músicos de reconocido prestigio.

Twitter permite enviar textos cortos con un máximo de 140 caracteres llamados tweets. Estos textos se muestran en la página principal del usuario que los ha generado y pueden ser vistos por el resto de usuarios de la aplicación. Lo podrán ver unos usuarios u otros dependiendo de las restricciones que haya marcado el autor.

Los usuarios pueden seguir a otros usuarios *"followers"* y tienen la opción de marcar los tweets como favoritos o incluso "*retweetearlos"* para compartir la información.

Twitter es muy popular en entornos de negocio y de actualidad. La mayor parte de los usuarios que usan Twitter son adultos mayores que no han utilizado otro sitio social con anterioridad.

Es una forma muy sencilla de ofrecer a un usuario reportar su opinión sobre cualquier tema de actualidad o de relevancia.

2.1.2 Facebook

Facebook es una herramienta social que pone en contacto a la gente con sus amigos y con otras personas que trabajan, estudian y viven en su entorno. Originalmente era un sitio para estudiantes de la Universidad de Harvard, pero se abrió a cualquier persona con una cuenta de correo electrónico.

Para tener contactos en Facebook es necesario aceptar una petición de amistad y así poder mostrar tu información a tus amigos. Facebook tiene alrededor de 900 millones de usuarios en todo el mundo.

Originalmente fue creada para el intercambio de fotos pero hoy en día tiene muchas más aplicaciones: compartir videos, mensajes, juegos, páginas publicitarias, etc.

Al igual que Twitter puedes compartir información de tus contactos y tiene la posibilidad de darle a *"me gusta"* y comentar toda la información.

También existe la posibilidad de crear grupos de gente con los mismos intereses.

2.1.3 LinkedIn

LinkedIn es la mayor red profesional del mundo creada en 2002 con 300 millones de usuarios en más de 200 países y territorios de todo el mundo.

Su misión es conectar a los profesionales del mundo para ayudarles a aumentar su productividad y rendimiento. Al unirte a LinkedIn se obtiene acceso a personas, empleos, noticias, actualizaciones e información que sirve de ayuda para destacar en el campo profesional que se desee.

En LinkedIn los usuarios registrados pueden tener una lista con información de contactos de las personas con quienes tienen algún nivel de relación, llamado Conexión. Esta lista de conexiones se pueden usar para encontrar puestos de trabajo y oportunidades de negocio recomendados por alguien de la red de contactos, subir el currículum vítae o diseñar el propio perfil con el fin de mostrar experiencias de trabajo y habilidades profesionales, etc.

Además, existen varias bases de datos públicas de donde obtener datos para su análisis. El más usado en los trabajos de *Data Mining* es el UCI Machine Learning Repository que contiene varios conjuntos de datos para probar los algoritmos. También se pueden extraer del API de Twitter o del API de Facebook.

Ilustración 1. Herramientas Extracción de Datos

2.1.4 UCI Machine Learning Repository

¿Qué ofrece y cuál es su utilidad?

La UCI Machine Learning Repository es un conjunto de bases de datos, teorías de dominio y generadores de datos que son utilizados por la comunidad de aprendizaje automático para el análisis empírico de los algoritmos de aprendizaje automático. El archivo fue creado en 1987 por David Aja y los estudiantes graduados compañeros en UC Irvine. Desde entonces, ha sido ampliamente utilizado por los estudiantes, educadores e investigadores de todo el mundo como una fuente primaria de los conjuntos de datos de la máquina de aprendizaje

Actualmente cuentan con 290 conjuntos de datos como un servicio a la comunidad de aprendizaje automático.

¿Cómo se accede?

Se accede a través de su página web: http://archive.ics.uci.edu/ml/index.html

¿Gratuito o de pago?

Es un recurso gratuito y ofrecen la posibilidad de enviar conjuntos de datos para almacenarlos en su web.

2.1.5 API de Facebook

¿Qué ofrece y cuál es su utilidad?

Facebook va más allá de su portal de interacción entre los usuarios, ofrece realmente una plataforma completa con herramientas para desarrolladores donde se pueden hacer aplicaciones para la Web, móviles y Facebook. Estas aplicaciones ponen en servicio un sin fin de métodos y propiedades para que nuestras APPs o Webs puedan convertirse en potentes sistemas de marketing, mediante la recolección de datos y comportamientos de los usuarios.

La API está desarrollada con una extensa compatibilidad a la mayoría de los SDK actuales disponibles, PHP, JavaScript, Android e IOS SDK, ActionScript etc. También está disponible una cantidad básica de plugins para la interacción de nuestros sitios con Facebook (Baraldi, 2012).

¿Cuáles son sus limitaciones?

El problema concreto que se presenta es que al ser tan extensa y poseer tantas posibilidades a la hora de listar ventajas o a elegir que lenguaje es el más adecuado se puede volver una tarea desalentadora.

¿Cómo se accede?

Para poder utilizar el API que Facebook proporciona a los desarrolladores es necesario registrarse como desarrollador en la web de Facebook. Para ello, sólo hay que inscribirse dentro de esta web http://developers.facebook.com/ y así obtener los datos necesarios para poder registrar una nueva aplicación que se pueda usar para conseguir datos de Facebook.

¿Gratuito o de pago?

Es una herramienta gratuita.

2.1.6 APIs de Twitter

¿Qué ofrece y cuál es su utilidad?

A través de las API de Twitter cualquiera puede crear aplicaciones que comuniquen con el servicio de la red social (Álvarez).

Twitter ofrece tres APIs: Streaming API, REST API y Search API aplicables a necesidades diferentes.

Streaming API

Para quién	Para todos los desarrolladores
Qué necesidad cubre	El Streaming API proporciona un *subset* de tweets en casi tiempo real
Cómo lo hace	Se establece una conexión permanente por usuario con los servidores de Twitter y mediante una petición http se recibe un flujo continuo de tweets en formato json
Beneficios	Se puede obtener una muestra aleatoria, un filtrado por palabras claves o por usuarios. Ofrece el perfil completo del autor en el momento de la escritura del tweet.
Aplicaciones	Twitter
Disponibilidad	Disponible actualmente

Ilustración 2. Streaming API Twitter

Rest API

Para quién	Para todos los desarrolladores
Qué necesidad cubre	Suministra los tweets con una profundidad en el tiempo de 7 días que se ajustan a la query solicitada
Cómo lo hace	Funciona por HTTP y se accede a partir de URLs
Beneficios	Es posible filtrar por, cliente utilizado, lenguaje y localización. No requiere autenticación y los tweets se obtienen en formato json o atom. Ofrece una información limitada del tweet (id, sreen_name y url del avatar)
Aplicaciones	Twitter
Disponibilidad	Disponible actualmente

Ilustración 3. Rest API Twitter

Search API

Para quién	Para todos los desarrolladores
Qué necesidad cubre	Ofrece a los desarrolladores el acceso al core de los datos de Twitter. Todas las operaciones que se pueden hacer vía web son posibles realizarlas desde el API. Soporta los formatos: xml, json, rss, atom
Cómo lo hace	Funciona por HTTP y se accede a partir de URLs
Beneficios	Acceso al core de los datos. Ofrece el perfil completo del autor en el momento de la escritura del tweet.
Aplicaciones	Twitter
Disponibilidad	Disponible actualmente

Ilustración 4. Search API Twitter

¿Cuáles son sus limitaciones?

El uso de las APIs de Twitter está limitado, por lo que las aplicaciones no pueden conectarse un número indeterminado de veces para realizar cualquier solicitud. Sin embargo, los límites serían más o menos aceptables para páginas personales y proyectos pequeños. En el caso que se desee construir sistemas que hagan un uso intensivo del API de Twitter, estaría la posibilidad de registrar la aplicación. Los límites de acceso al API sin registro son 150 solicitudes por hora, mientras que para aplicaciones registradas en la "whitelist" podrían llegarse a hacer 20.000 solicitudes por hora.

¿Cómo se accede?

Para poder utilizar el API de Twitter es necesario crear una aplicación en su página para tener las claves secretas con el acceso.

Se puede crear entrando a http://dev.twitter.com y registrándose. Después se pulsa sobre *"Mis aplicaciones"* y se rellena la información en la función *"Create an Aplication"*. Por último queda ingresar la nueva app y generar las claves.

¿Gratuito o de pago?

Es una herramienta gratuita.

Todas las herramientas comentadas en este punto suponen una fuente de información muy valiosa que podría utilizarse para comprobar, a tiempo real, lo que piensa la gente acerca de un tema específico. Una de los campos donde tiene más utilidad es en lo estudios de mercado.

La razón principal por la que se ha decidido utilizar Twitter como red social para la extracción de contenido es la cantidad de facilidades que ofrece a la hora de la obtención de información y que posee una estructura estándar con un límite de 140 caracteres que permite tratar el contenido de manera ágil y rápida.

2.1.7 Wrappers

¿Qué ofrece y cuál es su utilidad?

Se define de manera general como programas para la extracción de datos desde fuentes heterogéneas. La descripción que encontramos en JISBD 2003 (Alicante) es la siguiente: "*un sistema software que encapsula a las fuentes de datos. (…)Los wrappers realizan un proceso de transformación de las consultas recibidas en un conjunto de llamadas a los métodos de acceso a las fuentes de datos. Los resultados recuperados (…) son interpretados y transformados en documentos estructurados*" (Rodríguez Gil-Ortega).

Podemos clasificarlos en dos grupos según su generación:

- Generación Manual: Se escriben la reglas de extracción a mano por humanos examinando algunas páginas
- Wrappers de Inducción: El wrapper se contruye automáticamente aprendiendo de un conjunto de recursos tomados a modo de ejemplo. Pueden estar basados en

heurísticos (se leen reglas predefinidas) o basados en conocimiento (se aprenden las reglas)

María José Rodríguez Gil-Ortega (Profesora de la Universidad Politécnica de Madrid) describe una serie de propuestas de wrappers:

TSIMMIS: The Standfor-IBM Manager of Multiple Information Sources

Descripción	Nace como un proyecto enfocado principalmente al desarrollo de herramientas que faciliten una integración rápida de fuentes de información heterogéneas. Fueron de los primeros en pensar en el problema de la extracción de información. Está financiado por DARPA (Defense Advanced Research Projects Agency).
Limitaciones	No existe una visión descriptiva de las fuentes de datos subyacentes (el usuario tiene que saber a dónde va a preguntar), carece de ontologías para el enriquecimiento semántico y la generación de los wrappers se hace de forma manual.

Ilustración 5. TSIMMIS: The Standfor-IBM Manager of Multiple Information Sources

DISCO: Distributed Information Search Component

Descripción	Busca la integración de varias bases de datos al mismo tiempo basándose en el uso de mediadores. Estos mediadores pueden ser mediadores clásicos, que aceptan consultas y las transforman en consultas particulares que pueden ser distribuidas a las bases de datos integradas en el sistema o mediadores catálogo, que guardan información relativa a los elementos que componen el sistema. Está financiado por INRIA (Institute National de Recherche en Informatique et en Automatique, Francia).
Limitaciones	Al igual de TSIMMIS, no proporciona al usuario descripción semántica de los repositorios y en este caso no considera el problema de la integración de datos de fuentes heterogéneas.

Ilustración 6. DISCO: Distributed Information Search Component

INFORMATION MANIFOLD

Descripción	Fue desarrollado en los laboratorios AT&T Bell. El objetivo principal de IM es ofrecer un acceso uniforme a una colección de fuentes de información heterogéneas en la Web. Su arquitectura se sustenta en una base de conocimiento (denominada world-view) que contiene un modelo rico de dominio que permite la descripción de las propiedades de las fuentes de información. IM usa el modelo de datos objeto-relacional para representar el world-view.
Limitaciones	Como limitaciones cabe destacar que al estar la world-wide definida sobre el modelo relacional es muy rígido y no proporciona capacidad de inferencia, tampoco trata el problema de la integración semántica. A su favor, minimiza el número de repositorios accedidos.

Ilustración 7. Information Manifold

OBSERVER: Ontology Based System Enhanced with Relationships for Vocabulary Heterogeneity Resolution.

Descripción	Es un sistema para la extracción de información de distintos repositorios haciendo uso de ontologías definidas por expertos y responsables de los datos. Permite que las ontologías y los repositorios sean desarrollados de forma independiente. Algunas características a su favor son que es uno de los sistemas más completos de este tipo y han tenido en cuenta muchas limitaciones como la pérdida de información.
Limitaciones	No proporciona integración de ontologías de forma automática y si un repositorio se *casa* con una ontología ya no se puede casar con ninguna más.

Ilustración 8. OBSERVER: Ontology Based System Enhanced with Relationships for Vocabulary Heterogeneity Resolution.

¿Cómo se accede?

Son generados de manera autónoma por lo que el acceso es directo a cada persona.

¿Cuáles son sus limitaciones?

Las limitaciones dependen de las funcionalidades que el autor de cada wrapper añada en él. Para que la realización de estos wrappers sea menos costosa se buscan metodologías para desarrollar wrappers de forma semi-automática.

¿Gratuito o de pago?

Es un recurso gratuito, únicamente necesitas un programa de generación de código.

2.2 DATA MINING

El *Data Mining* se define como el proceso de descubrimiento de patrones en los datos. El patrón descubierto debe ser significativo y proporcionar una ventaja, por lo general económica. Los datos son presentados en grandes cantidades. (Witten, 2005)

Otra definición de *Data Mining* es la siguiente: "*el proceso de analizar datos desde diferentes perspectivas con el objetivo de resumir los datos en segmentos de información útiles.* "

El *Data Mining* surge con el objetivo de ayudar a comprender el contenido de los repositorios de datos mediante estadísticas, algoritmos de inteligencia artificial, redes neuronales, etc... Los datos no son más que un conjunto de materia prima bruta a los que una persona les otorga un significado y los convierte en información. Al realizar las funciones de análisis de esa información se genera el conocimiento. Esta información puede ser usada para incrementar beneficios, reducir costes, etc.

Los casos relacionados con el *Data Mining* siguen un proceso común que está formado por cuatro etapas:

1. **Determinación de los objetivos.** Se delimitan los objetivos que se desean conseguir.
2. **Preprocesamiento de los datos.** Se realiza un proceso de selección, limpieza, enriquecimiento, reducción y transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de Data Mining.
3. **Determinación del modelo.** Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.
4. **Análisis de los resultados.** Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El interesado determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

Un usuario de *Data Mining* realiza este trabajo con la finalidad de encontrar una de estas cuatro relaciones (Marín Diazaraque, 2009):

1. **Clases:** Con la creación de clases se asigna una serie de datos a un grupo prefijado con la finalidad de evitar realizar una clasificación errónea de datos.
2. **Clusters:** se construyen grupos de observaciones similares según un criterio prefijado. El proceso se clustering consiste en subdividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo más cercano posible a otro elemento y los grupos diferentes estén lo más lejos posible uno de otro.
3. **Asociaciones:** son usadas para identificar agrupaciones entre variables.
4. **Patrones Secuenciales:** se intenta identificar un conjunto de patrones de comportamiento y tendencias.

En la actualidad el uso del *Data Mining* está extendido por todos los campos. En uno de los que tiene más presencia es en los negocios, esto es debido a la necesidad que sienten las empresas de conocer los hábitos y los gustos de los consumidores para poder realizar mejores campañas de marketing.

2.3 PREPROCESAMIENTO DE LOS DATOS

La función de este proceso es la de eliminar los datos que no aporten información relevante con el objetivo de simplificar el contenido global.

Pyle en su publicación de 1999 define el preprocesado de datos como: *"El propósito fundamental de la preparación de los datos es la manipulación y transformación de los datos sin refinar para que la información contenida en el conjunto de datos pueda ser descubierta o estar accesible de forma más fácil"* (Pyle, 1999). Los datos obtenidos a partir de las herramientas que se han comentado anteriormente pueden ser incompletos, ruidosos y/o inconsistentes. Las acciones que se realizan en este módulo son (León Guzmán, 2010):

Limpieza

- Resuelve redundancias
- Chequea y resuelve problemas de ruido, valores perdidos, elimina outliers
- Resuelve inconsistencias/conflictos entre datos

Integración

- Obtiene los datos de diferentes fuentes de Información
- Resuelve problemas de representación y codificación
- Integra los datos desde diferentes tablas para crear información homogénea

Transformación

- Los datos son transformados o consolidados de forma apropiada para la extracción de información. Diferentes vías:
    - Resumen de datos
    - Operaciones de agregación, etc.

Reducción

- Discretización
- Selección de Instancias (objetos)
- Selección de características

Esto que se ha contado es una pequeña parte de una gran labor que es el preprocesado de datos con un montón de acciones más por describir. A continuación se indican dos referencias dónde se puede encontrar más información acerca de esta función: (Botía, 2009) y (Cardenas Montes, 2013)

Un inconveniente del preprocesado de datos es que el siguiente: éste no es un área totalmente estructurado con una metodología concreta de actuación para todos los problemas. Cada problema puede requerir una actuación diferente, utilizando diferentes herramientas de preprocesamiento.

En este proyecto se han creado distintos métodos de manera autónoma para el tratamiento de esta información como se podrá ver en la sección (Ver Apartado 4. Diseño e Implementación del sistema)

2.4 HERRAMIENTAS DE APOYO PARA EL ANÁLISIS DE DATOS

Una vez que se han preprocesado los datos y se ha obtenido la información relevante, ésta debe ser analizara para generar valor.

En este apartado se describirán las diferentes herramientas que sirven de ayuda para realizar este proceso.

Ilustración 9. Herramientas Análisis de Datos

La mayoría de las Redes Sociales tienen sus propias APIs con la finalidad de ofrecer una plataforma completa con herramientas para los desarrolladores que quieran hacer aplicaciones móviles/web o bien para la recolección de los datos y comportamientos de sus usuarios. Este es el caso de Facebook, Twitter, Google, etc. Pero también podemos encontrar APIs ajenas a estas redes que realizan la misma función abarcando más de una de estas aplicaciones.

Es importante resaltar que en muchos países del mundo es un tema que lleva mucho tiempo en auge, pero en España aún no ha acabado de emerger, por lo que la mayoría de las herramientas se basan fundamentalmente en el diccionario inglés, debido principalmente a que el procesamiento del lenguaje es mucho más sencillo en este idioma.

Para este trabajo se ha recogido una pequeña muestra que será clasificada (Alonso i Alemany, 2005) y descrita a continuación.

2.4.1 Análisis Sintáctico

Mediante el análisis sintáctico se pueden crear grupos lingüísticos

2.4.1.1 FreeLing

¿Qué ofrece y cuál es su utilidad?

FreeLing es una librería gratuita orientada a la prestación de servicios de análisis de lenguaje. Está desarrollado por el TALP de la UPC y liderado en la actualidad por Lluis Padró. Con esta herramienta los desarrolladores pueden utilizar los recursos lingüísticos por defecto (diccionarios, lexicones, gramáticas, etc), ampliarlos, adaptarlos a dominios particulares, o desarrollar otros nuevos para necesidades especiales de las aplicaciones (Padró L.).

El proyecto se estructura como una librería que puede ser llamada desde cualquier aplicación de usuario que requiera servicios de análisis del lenguaje. El software se distribuye como código abierto bajo una licencia GNU General Public License y bajo licencia dual a empresas que deseen incluirlo en sus productos comerciales.

Está concebido con la idea de que se puedan desarrollar potentes aplicaciones de PLN, y orientado a facilitar la integración con las aplicaciones de niveles superiores de los servicios lingüísticos que ofrece.

Su arquitectura se basa en un enfoque de dos capas cliente-servidor: una capa básica de servicios de análisis lingüístico y una capa de aplicación que, actuando como cliente, realiza las peticiones deseadas a los analizadores y usa su respuesta según la finalidad de la aplicación.

La arquitectura interna de la librería se estructura según dos tipos de objetos: los que almacenan datos lingüísticos con los análisis obtenidos y los que realizan el procesamiento en sí.

Los resultados de FreeLing son un análisis lingüístico en una estructura de datos, cada aplicación de usuario final puede tener acceso a esos datos y procesarlos según sea necesario.

El paquete de FreeLing ofrece un programa de aplicación bastante completo que permite a un usuario final sin conocimientos de programación, obtener el análisis de un texto.

En la actualidad se está ofreciendo la versión 3.1 que ofrece las siguientes novedades:

- Freeling 3.1 es segura para subprocesos lo que permite el procesamiento en paralelo (por ejemplo, en servicios web).
- Ampliada API para módulos de análisis.
- Mejora de la API para administrar y navegar resultante estructuras lingüísticas.
- Nuevas funcionalidades para hacer frente al texto no estándar. SED- Búsqueda basada en palabras del diccionario similares a la dada, con base en FOMA.
- Módulo de expresión regular que permite elegir entre boost :: regex y boost :: xpressive como motor subyacente.
- Nuevos idiomas: Hasta el punto de venta/etiquetado para el francés. Soporte parcial para Checa y Eslovenia.
- Instalación más sencilla : menos dependencias externas, sólo desde fuera de la caja de paquetes libboost
- Más fácil de compilación en MacOSX y Windows (con MSVC).

Los lenguajes soportados actualmente por la herramienta son: asturiano, catalán, inglés, francés, gallego, italiano, portugués, ruso, esloveno, español, y galés.

¿Cuáles son sus limitaciones?

Hay que tener en cuenta que FreeLing no es una herramienta de análisis de texto orientado al usuario; Es decir, no está diseñado para ser fácil de usar, ni se obtienen los resultados en un formato determinado. (Padró L. , Octubre 2013.)

¿Cómo se accede?

Puede descargarse el programa directamente desde su web http://nlp.lsi.upc.edu/freeling/ donde además se encuentran las instrucciones de descarga, un manual de usuario, una demo de la aplicación, etc.

¿Gratuito o de pago?

Es un recurso gratuito.

Al igual que ha ocurrido con WordNet no ha sido posible instalarlo en mi proyecto por motivos de compatibilidad.

2.4.2 Análisis Morfológico

Las herramientas descritas en esta sección se encargarán de definir la categoría de cada una de las palabras analizadas. Todos los analizadores morfológicos y sintácticos tienen un diccionario, en los casos de analizadores de código abierto, el diccionario es accesible. Además también se pueden encontrar en esta sección correctores ortográficos.

2.4.2.1 Normalizador RAE

Por último se define una herramienta proporcionada por la RAE. Se trata de un normalizador de texto. El propósito del normalizador es reducir o eliminar la potencial redundancia de datos y dependencias incoherentes dentro de un texto. El funcionamiento de este sistema se explica con un ejemplo en el apartado (4.6.3. Normalizador RAE)

2.4.2.2 Apicultur

¿Qué ofrece y cuál es su utilidad?

Apicultur se define como una plataforma de APIs y ha sido desarrollada por *Molino de Ideas* con la solución Api Manager de WSO2. En ella se pueden encontrar APIs lingüísticas de lematización, fonética, análisis morfológico, palabras incluidas en el diccionario español, etc. Todas estas funciones pueden ser incorporadas a otras aplicaciones gracias a la API. (Sánchez Suarez)

Los idiomas soportados por esta herramienta son inglés y español.

En la Tienda de Apicultur, están disponibles las distintas APIs y se puede navegar por ellas según distintas categorías. Al pinchar sobre ellas se puede ver una breve descripción sobre su funcionamiento, la URL de acceso y los datos del propietario, así como los comentarios que otros usuarios han dejado sobre la API y la valoración que han hecho. En la pestaña Documentación se puede consultar toda la información necesaria para utilizar la API.

¿Cuáles son sus limitaciones?

No hay limitaciones definidas de la herramienta. En cuanto a las limitaciones tanto de velocidad como de capacidad se resuelven aumentando la categoría, y por tanto viene asociado un aumento de coste.

¿Cómo se accede?

Para poder acceder y usar una API, es necesario darse de alta en la Tienda de Apicultur https://store.apicultur.com/.

Lo primero que hay que hacer es crear una aplicación. Una vez identificado, desde MySubscriptions se puede crear una nueva aplicación o bien usar la que viene por defecto. Una aplicación no es más que una manera de almacenar las suscripciones a las APIs de una manera más ordenada. El objetivo de las aplicaciones es facilitar la gestión y organización de los suscripciones a las APIs, ya que todas las APIs suscritas bajo una misma aplicación tendrán la misma clave.

Una vez creada la aplicación hay que seleccionar nivel y suscribirse a la API. A continuación, entrar en la API que se quiera utilizar. Antes de suscribirse a una API se puede probar su funcionamiento con la documentación swagger. En el perfil de cada API hay que pinchar sobre el método e introducir diversos parámetros para comprobar qué devuelve la API. Para suscribirse a la API necesitas seleccionar en el menú desplegable bajo qué aplicación se quiere guardar la suscripción a esa API. Selecciona también el nivel de suscripción. Hay tres modalidades (oro, plata, bronce), atendiendo al número máximo de consultas que se pueden hacer a la API por minuto. Por último hay que pinchar en el botón "Suscribe" para suscribirte a la API.

Una vez suscrito, en el mismo apartado se encontrará además la URL que hay que usar para acceder a la AP la información sobre el propietario de la API en cuestión. En la pestaña Documentation se puede acceder a la documentación de la API.

Para su uso será necesario obtener la autorización. En el apartado MySubscription podrán verse las suscripciones clasificadas por aplicación. Para cada aplicación es necesario generar las claves que permitirán consultar las APIs correspondientes. Para generarlas, hay que pinchar en "Generate" en el apartado "Production". La clave "Access token" es la necesaria para poder acceder a la API.

¿Gratuito o de pago?

Apicultur dispone de diferentes tarifas dependiendo del tamaño de la clase. Además te ofrecen 20€ mensuales de manera gratuita y una serie de garantías definidas en su página web.

Apicultur es una de las herramientas utilizadas en este proyecto y cuando hablé con ellos me ofrecieron todo tipo de facilidades a la hora de realizar el proyecto, comentándome que al ser un proyecto de investigación no tendría que pagar nada por utilizarlo.

2.4.3 Análisis Semántico

2.4.3.1 WordNet

¿Qué ofrece y cuál es su utilidad?

WordNet es una base de datos léxica. En versiones anteriores agrupaba palabras solamente en inglés en conjuntos de sinónimos llamados synsets, proporcionando definiciones cortas y generales, y almacenaba las relaciones semánticas entre los conjuntos de sinónimos. Su propósito es doble: producir una combinación de diccionario y tesauro cuyo uso sea más

intuitivo, y soportar análisis automático de texto y aplicaciones de Inteligencia Artificial (Troyano) (A. Montejo-Ráez).

Los synsets son conjuntos de palabras sinónimas que se presentan relacionadas entre sí a través de la hiperonimia, por lo que el resultado final es una red semántica. Para cada synset se aporta la definición, compartida por los diferentes miembros del grupo de sinónimos, y, en algunos casos, se presentan también ejemplos de uso de algunos de ellos.

Para la creación del WordNet 3.0 (Page.), se ha partido del recurso en inglés para adaptarlo al español. Algunos de los datos contenidos en este léxico, como las etiquetas de la anotación morfosintáctica y semántica, han sido extraídos de dicho recurso e incorporados directamente al léxico español y otros, como las palabras de las entradas, las definiciones y los ejemplos, han sido traducidos manteniendo un enlace con el original, con el fin de obtener un corpus paralelo inglés-español.

Actualmente el proyecto ha finalizado y se han traducido aproximadamente unas 15.000 glosas, lo cual quiere decir que están disponibles para el español aproximadamente unas 30.000 entradas léxicas (nominales y verbales).

El impacto de dicha herramienta ha sido espectacular en el campo del procesamiento del lenguaje natural (PLN), donde se utiliza como un estándar para la desambiguación semántica automática a nivel de palabra.

A partir de WordNet se construye otro recurso léxico llamado SentiWordNet (Definido más adelante)

¿Cuáles son sus limitaciones?

A diferencia de otros diccionarios, WordNet no incluye información sobre la etimología, pronunciación y la forma de los verbos irregulares y contiene solo información limitada sobre su uso. La información lexicográfica y semántica actual se mantiene en lexicographical files, que son procesados por una herramienta llamada grind para producir la base de datos distribuida. Ambos, el grind y los lexicographer files, están disponibles libremente en una distribución separada, pero la modificación y mantenimiento de la base de datos requiere experiencia. A pesar de que WordNet contiene un rango suficientemente amplio de palabras comunes, no cubre vocabulario de un dominio específico. Como está diseñada en primer lugar para actuar como capa subyacente para diferentes aplicaciones, esas aplicaciones no pueden ser usadas en dominios específicos que no son cubiertos por WordNet (Fernández Montraveta).

La versión en español no está desarrollada por completo y en este proyecto se han obtenido una serie de fallos que no han hecho posible que pudiera utilizarla. Además no es compatible aún con Windows ni con Linux de 64 bits.

¿Cómo se accede?

Este recurso está actualmente disponible en la red http://wordnet.princeton.edu/wordnet/download/ .

La novedad que presenta la versión 3.0 es que el corpus que constituyen las definiciones y los ejemplos está etiquetado a nivel morfosintáctico y semántico a nivel de palabra.

¿Gratuito o de pago?

Se puede consultar y utilizar de manera gratuita con fines de investigación.

2.4.4 Análisis del Sentimiento Completo

Esta herramienta se encarga del análisis completo del sentimiento de la información obtenida en internet.

2.4.4.1 APIs de Textalytics

¿Qué ofrece y cuál es su utilidad?

Las tres APIs que se van a definir a continuación comparten varias características por lo que la descripción será conjunta. Las características específicas de cada una se mostrarán en una tabla.

Textalytics es fruto de (Daedalus, Textalytics) *y* ofrece APIs estándar de alto nivel para diversos sectores y escenarios de uso, liberando así a los usuarios de la complejidad técnica de estos desarrollos y reduciendo su "*time-to-market*". Se comercializa en modo SaaS (Software as a Solutions).

Daedalus es una empresa española especializada en el procesamiento de contenido no estructurado, procesamiento de lenguaje natural y minería de texto. *Daedalus* resuelve las "3 Vs" en análisis de contenido no estructurado: variedad (trata contenidos de cualquier canal), velocidad (análisis en tiempo real) y volumen (tecnología multiproceso escalable y disponibilidad en la nube).

Las principales ventajas de Textalytics son:

- Ayuda a los usuarios a centrarse en sus aplicaciones y su negocio, no en la tecnología. Sus servicios de alto nivel y orientados al negocio ocultan la dificultad técnica y salvan la brecha entre la tecnología semántica y las necesidades de negocio.
- Además de su funcionalidad relevante para el negocio, los ejemplos de código, plug-ins y SDK incluidos en el producto proporcionan a los desarrolladores una rápida curva de aprendizaje, una usabilidad superior y un time-to-market más corto.
- Al proporcionarse como servicio en la nube con un modelo freemium, Textalytics constituye una oferta de alto valor, bajo riesgo y fácil consumo.
- Incorpora, entre otras, una API para Análisis de Medios (tanto sociales como tradicionales), una API de Publicación Semántica (para medios de comunicación y editoriales)… y una Core API (para usuarios que necesitan una funcionalidad más desagregada y personalizable).

- Es multilingüe: español, inglés, francés, italiano, portugués y catalán.

Las características específicas de cada una de las APIs mencionadas anteriormente son (De Pablo, Octubre 2013) :

Análisis de Medios

Para quién	Agencias y departamentos de marketin/comunicación/seguimiento de medios, etc.
Qué necesidad cubre	Entender lo que se dice en los medios sociales y tradicionales en volumen, velocidad y variedad.
Cómo lo hace	Servicios personalizables para monitorización de marcas, organizaciones, personas, temas, análisis de sentimiento, etc.
Beneficios	Información más precisa, completa y "actuable" de todo tipo de medios, en tiempo real y sin importar el volumen.
Aplicaciones	Seguimiento de medios, análisis competitivo, social TV, publicidad enfocada.
Disponibilidad	Disponible actualmente.

Ilustración 10. API Textalytics, Análisis de Medios

Publicación Semántica

Para quién	Medios de comunicación (prensa, radio, TV), editoriales, publicadores de contenido.
Qué necesidad cubre	Producir contenidos más valiosos, más rápidamente y con menor coste, monetizarlos mejor.
Cómo lo hace	Servicios personalizables de etiquetado, enriquecimiento, revisión.
Beneficios	Mayores posibilidades de caracterizas, descubrir, encontrar, reutilizar, modularizar, relacionar, combinar, personalizar… contenidos.
Aplicaciones	Publicación semántica dinámica, productos a medida, gestión de archivos/activos digitales, publicidad enfocada (contexto).
Disponibilidad	Disponible actualmente.

Ilustración 11. API Textalytics, Publicación Semántica

Core

Para quién	Desarrolladores, integradores en cualquier sector/aplicación.
Qué necesidad cubre	Desarrollar a medida sus propios pipelines de procesamiento semántico.
Cómo lo hace	Servicios semánticos horizontales y granulares: topics, sentimiento, clasificación, perfilado demográfico, etc.
Beneficios	Flexibilidad y personalización máximas.
Aplicaciones	General.
Disponibilidad	Disponible actualmente.

Ilustración 12. API Textalytcis. Core

¿Cuáles son sus limitaciones?

A continuación se describen las limitaciones que se pueden encontrar en esta herramienta y las recomendaciones que proporciona *Daedalus* para evitarlas o intentar que afecten lo mínimo posible:

Número de llamadas: El crédito en las API de Textalytics se cuenta en un número de créditos. En el caso de la API de Análisis de Medios, cada crédito equivale a una palabra procesada. De esta forma no importa si los documentos son grandes o pequeños, sino la cantidad de texto que se procesa. Si se dispone de una licencia gratuita tienes hasta 200 000 créditos al mes. En *Daedalus* creen que con este límite se puede estar seguro de que los resultados son los más adecuados para la aplicación. Además, a efectos de conteo solo se tienen en cuenta las llamadas correctas que se han ejecutado de forma completa.

Tamaño del documento: El tamaño máximo de la petición tiene un límite de 1MB. Si los documentos superan este tamaño en *Daedalus* recomiendan que se hagan divisiones naturales como capítulos y se envíe por partes.

Tasa de llamadas: El plan gratuito está limitado a 5 llamadas por segundo y por licencia. Si la aplicación excede este número de respuestas puede recibir un error por parte del servicio (HTTP 429, error interno 104). Se pueden evitar estas peticiones erróneas si se implementa un tiempo de espera razonable entre llamadas. En caso de que sea necesario procesar contenido a una tasa mayor se puede contratar un plan superior y hacer un acuerdo de servicio (SLA) a medida.

¿Cómo se accede?

Se puede adquirir la herramienta en su página web http://textalytics.com .

¿Gratuito o de pago?

Tiene varios modelos entre los que se puede encontrar uno gratuito, disponible mediante el registro en la página web. Los otros tres modelos (professional, business y enterprise) sí son de pago y en el apartado de precios de la página web vienen descritas sus características.

2.4.4.1.1 Sentimentalytics

En este apartado se va a hacer referencia a una funcionalidad especial dentro de la herramienta de Textalytics.

¿Qué ofrece y cuál es su utilidad?

Sentimentalytics es una herramienta cuya funcionalidad es el análisis de sentimiento y monitorización de medios sociales (redes sociales, blogs, microblogs, sitios de noticias) escaneando e identificando automáticamente lo más importante.

Sentimentalytics aplica tecnologías semánticas para escanear y etiquetar automáticamente los tweets, posts, y noticias mostrados en los diferentes timelines, canales, consultas, etc. que aparecen en tu navegador.

El etiquetado con entidades/polaridades/categorías permite identificar aquellas menciones, opiniones y temas que merecen un tratamiento individual, liberando a Community Managers y Agencias de la lectura rutinaria de posts irrelevantes y habilitando la respuesta en tiempo real.

Las ventajas del uso de esta herramienta son:

- La tecnología semántica más avanzada para el análisis de medios sociales.
- Analiza los medios sociales más habituales y en varios idiomas.
- Escanea e identifica lo más relevante.
- Una nueva forma de ver y navegar por tus timelines.
- Exhaustivo y en tiempo real.
- Fácil e intuitivo.
- Un simple plug-in compatible con tu browser.

¿Cuáles son sus limitaciones?

Una limitación que tuvo esta herramienta al comienzo fue la indisponibilidad de exportar los datos que se generaban a un programa como puede ser Excel, pero ha sido actualizado hace poco. Además se ha añadido el Francés a los idiomas soportados.

¿Cómo se accede?

Se accede mediante el registro en su página web y la descarga de un plugin. Además en el apartado de soporte de su página describen las especificaciones técnicas necesarias para usar la aplicación. https://sentimentalytics.com/soporte-analisis-sentimiento-semantico

¿Gratuito o de pago?

Sentimentalytics ofrece gratis su funcionalidad de análisis de sentimiento y monitorización de medios sociales dentro de ciertos volúmenes de uso. Si se desea una funcionalidad más a medida dan su dirección de correo electrónica para ponerse en contacto con ellos.

2.4.4.2 SentiWordNet

¿Qué ofrece y cuál es su utilidad?

Tal y como se ha comentado anteriormente SentiWordNet se crea a partir de WordNet. Es un diccionario que proporciona información acerca de la orientación semántica en cuanto a emoción de los synsets. SentiWordNet devuelve, para cada synset, una tripleta de tres valores que miden la carga de “positividad”, “negatividad” u “objetividad” del mismo. La última versión (3.0) se ha generado a partir de las anotaciones manuales de versiones previas, propagando sobre el grafo dichos valores de emoción mediante un algoritmo de tipo random walk.

¿Cuáles son sus limitaciones?

Como es una herramienta creada a partir de WordNet, tiene sus mismas limitaciones.

¿Cómo se accede?

Se accede a él mediante su página web http://sentiwordnet.isti.cnr.it/ y a través de WorNet.

¿Gratuito o de pago?

Se puede usar de manera gratuita con fines de investigación y siempre que se mencione y se atribuya a sus autores.

2.4.4.3 Profiler Plus

¿Qué ofrece y cuál es su utilidad?

Profiler Plus es una herramienta de Social Science Automation junto con Ravenbrook, Ltd y tiene como cartera de clientes el gobierno, los negocios y el mundo académico. Estos se benefician de la velocidad y la precisión del análisis del texto automático. Mediante el uso de esta plataforma y la aplicación de diversos esquemas de codificación, es posible responder a las preguntas del mundo real para su negocio o industria. Con esta herramienta es posible realizar lo siguiente (ProfilerPlus):

- Seguimiento de los sentimientos en los medios sociales o en las marcas, empresas, personas, etc.
- Tomar decisiones plenamente informado sobre nuevos talentos para su equipo.
- Uso del análisis del lenguaje para evaluar posibles amenazas.
- Mitigar los riesgos mediante el seguimiento y la evaluación de datos.
- Comprender diferentes culturas y hacer comparaciones en varios idiomas acerca de estas.
- Perfil de los líderes de los países, empresas o cualquier otra persona en posiciones de poder.

Profiler Plus está disponible como un programa de escritorio de Windows con una interfaz gráfica de usuario, el procesador por lotes, componente de Windows, o el servicio web.

- El sistema proporciona control y transparencia en el proceso completo para permitir un proceso de análisis personalizado e individualizado.
- Existen modelos para múltiples idiomas: inglés, árabe, español, ruso y chino.
- La salida de datos se suministra en formatos flexibles y universales: TXT, XML, CSV

Esta tecnología y software posterior surgió de la investigación y la aplicación de diversos organismos gubernamentales. Después de diez años de desarrollo y apoyo para el gobierno, Profiler Plus ahora también se está utilizando para apoyar las aplicaciones más amplias en las áreas de análisis de medios de comunicación, evaluación de los medios y psicolingüística forenses. Profiler Plus está implementado en Common Lisp y, a diferencia de la mayoría de los sistemas basados en la estadística de PLN, Profiler Plus está enteramente basada en reglas que

proporcionan control y transparencia del proceso completo. Los resultados son interpretados por nuestros expertos analistas y / o importados en este tipo de programas de análisis de datos como Microsoft Excel, SPSS y SAS.

¿Cuáles son sus limitaciones?

Al ser una herramienta utilizada por el gobierno, una de las limitaciones que encontradas es el idioma. Con el fin de hacer operativo el *Profiler Plus* para el uso en cualquier idioma, se debe trabajar con un programa de edición denominado Xeditor, esta herramienta permite crear y editar todas las reglas de codificación que luego serán reconocidas por *Profiler Plus.*

¿Cómo se accede?

Para ver su funcionamiento es posible registrarse en la página web http://profilerplus.org

¿Gratuito o de pago?

Es un producto de pago y se comercializa en lotes de 100 test.

Es importante resaltar que todas las herramientas descritas en este apartado son sólo una pequeña muestra de una gran cantidad de herramientas que sirven para este tipo de procesado y análisis de datos. Además, las herramientas descritas aquí son las que realizan este tipo de acciones con datos escritos en español, como he comentado en la introducción en España aún no está tan desarrollado como en otros países donde hay una cantidad de herramientas mayor y también con una precisión mayor. En España actualmente se trabaja mucho en este entorno pero aún queda mucho trabajo por hacer.

2.5 VALIDACIÓN DEL ANÁLISIS REALIZADO

Para la validación de los resultados obtenidos a partir del motor creado se han utilizado las métricas de Precisión y Recall (Marrero, 2013).

Con la Precisión se mide la cantidad de documentos recuperados que son relevantes (Mide el ruido).

El Recall mide la cantidad de documentos relevantes que son recuperados (Mide la exhaustividad).

Estas dos métricas suelen tener una relación inversa. Si uno mejora el otro empeora. La preferencia de mejorar en uno o entro varía según la tarea y el modelo de usuario. Los motores web, por ejemplo, prefieren ofrecer precisión pero un médico o un abogado prefiere un recall alto.

Para definir las fórmulas utilizadas en estas métricas se hace necesario describir previamente cuatro conceptos más:

- **True Positives** (TP): son instancias pertenecientes a la clase que se clasifican correctamente en esa clase.
- **True Negatives** (TN): son instancias no pertenecientes a la clase y que no se clasifican como esa clase.
- **False Positives** (FP): son instancias no pertenecientes a la clase pero que se clasifican como esa clase.
- **False Negatives** (FN): son instancias pertenecientes a la clase pero que no se clasifican como esa clase.

Teniendo claros estos conceptos pasamos a definir el algoritmo utilizado para la validación:

- **Precision para la clase C:** es un valor entre 0 y 1. Su valor aumenta cuando hay pocos falsos positivos. Mide que las instancias clasificadas como clase C sean realmente de la clase C, aunque haya instancias de la clase C que se clasifiquen como otra clase

$$P(C) = TP / (TP + FP)$$

De esta forma, cuanto más se acerque el valor de la precisión al valor nulo, mayor será el número de documentos recuperados que no consideren relevantes. Si por el contrario, el valor de la precisión es igual a uno, se entenderá que todos los documentos recuperados son relevantes. Esta forma de entender la precisión introduce el concepto de ruido informativo y de silencio informativo.

- **Recall para la clase C:** es un valor entre 0 y 1. Su valor aumenta cuando hay pocos falsos negativos. Mide que las instancias de la clase C se clasifiquen como clase C, aunque otras instancias también se clasifiquen como clase C sin serlo.

$$R(C) = TP / (TP + FN)$$

Si el resultado de esta fórmula arroja como valor 1, se tendrá la exhaustividad máxima posible, y esto viene a indicar que se ha encontrado todo documento relevante que residía en la base de datos, por lo tanto no se tendrá ni ruido, ni silencio informativo: siendo la recuperación de documentos entendida como perfecta. Por el contrario en el caso que el valor de la exhaustividad sea igual a cero, se tiene que los documentos obtenidos no poseen relevancia alguna.

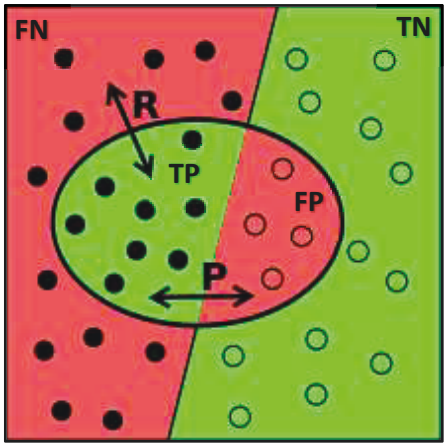


Ilustración 13. Precisión y Recall

De la agrupación de estas dos medidas surge F-Measure. Su algoritmo sigue la siguiente estructura:

F(C) = 2 * Precisión * Recall / (Precisión + Recall)

2.6 HERRAMIENTAS DE ANÁLISIS DE LA INFORMACIÓN OBTENIDA

Una vez realizado el análisis de los datos y comprobada su validez existe una serie de herramientas que permiten realizar un análisis de la información obtenida mediante gráficos, comparativas, etc.

2.6.1 Weka

Una de las herramientas que he utilizado en la carrera y que proporciona dicha información es Weka (Waikato)**.** Es mucho más completa ya que mediante esta herramienta se puede obtener información de la mayoría de los procesos descritos anteriormente. Es una colección de algoritmos de aprendizaje automático para tareas de *Data Mining*. Contiene herramientas para el pre-procesado de datos, clasificación, regresión, clustering, reglas de asociación y visualización. Es software de código abierto bajo la GNU General Public License.

Los puntos fuertes de Weka son:

- Está disponible libremente bajo la licencia pública general de GNU.
- Es muy portable porque está completamente implementado en Java y puede correr en casi cualquier plataforma.
- Contiene una extensa colección de técnicas para preprocesamiento de datos y modelado.
- Es fácil de utilizar por un principiante gracias a su interfaz gráfica de usuario.

Ilustración 14. Weka

2.6.2 Matlab

Otra de las herramientas utilizadas a lo largo de la carrera y que permiten realizar un estudio gráfico es Matlab. En su página Web (Moler) se define como un lenguaje de alto nivel y un entorno interactivo para el cálculo numérico, la visualización y la programación. Mediante MATLAB, es posible analizar datos, desarrollar algoritmos y crear modelos o aplicaciones. El lenguaje, las herramientas y las funciones matemáticas incorporadas permiten explorar diversos enfoques y llegar a una solución antes que con hojas de cálculo o lenguajes de programación tradicionales, como pueden ser C/C++ o Java™.

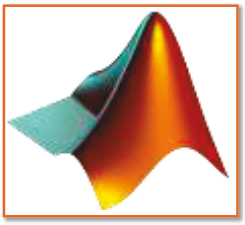


Ilustración 15. Matlab

2.6.3 Gephi

Gephi es un software de visualización y análisis de grafos orientado a todo tipo de redes y sistemas complejos y grafos dinámicos y jerárquicos. Dispone de varios algoritmos implementados para el análisis de Redes Sociales, como PageRank, HITS, etc.También presenta algoritmos para mejorar la visualización del grafo, utilizando diferentes diseños y es capaz de calcular diferentes métricas del grafo como su grado, intermediación, cercanía, densidad, longitud de la trayectoria, diámetro, modularidad o coeficiente de clustering.

Ilustración 16. Gephi

Se ejecuta en Windows, Linux y Mac OS X. Gephi es de código abierto y libre.

Tal y como se ha ido recalcando a lo largo de toda la memoria, estas son solo algunas de las herramientas que pueden utilizarse para analizar la información obtenida en los análisis de datos. Existen diversidad de herramientas de *Business Inteligence* y cada una con distintas características que se adaptan a los distintos entornos de la red. Empresas como IBM, Oracle, etc. están avanzando mucho en las herramientas destinadas a estas funciones.

Como dato de interés, en una noticia muy reciente (Computerworld, Junio 2014) se puede leer lo siguiente: *"El acuerdo entre IBM y Genesys permitirá que se combinen las capacidades de Watson Engagement Advisor con la "plataforma de experiencia del cliente" de Genesys. Esto permitirá mejorar la relación de las organizaciones con los consumidores finales, así como el servicio que ofrecen desde los centros de asistencia (contact centers) y los servicios de atención al cliente. Ambas ayudarán y asistirán a los teleoperadores mediante respuestas rápidas, apoyándose en el análisis de grandes volúmenes de datos. Con un clic en la solución Pregunta a Watson (Ask Watson)puede ayudar rápidamente a abordar y solucionar los problemas de los clientes y darles una respuesta para orientar sus decisiones de compra."* (Computerworld, Junio 2014).

2.7 UNA REFERENCIA ESPECIAL: SEPLN

No podría finalizar este apartado sin hablar del taller organizado por La Sociedad Española para el Procesamiento del Lenguaje Natural[1]acerca del análisis de sentimientos, denominado TASS (SEPLN, TASS - Workshop on Sentiment Analysis at SEPLN) (SEPLN, 2013). Con este taller se quiere fomentar la investigación en el campo del análisis de sentimiento en los medios sociales, especialmente en el idioma español. El principal objetivo es promover el diseño de nuevas técnicas y algoritmos y la aplicación de los ya existentes para la implementación de complejos sistemas capaces de realizar un análisis de sentimientos basados en opiniones de textos cortos extraídos de medios sociales (concretamente Twitter).

Una vez finalizado el congreso se realiza un documento que agrupa las actas que se han realizado para el evento. En este documento (Díaz Esteban, Alegria Loinaz, & Villena Román,

[1] Home Page: http://www.sepln.org/

2013) se indica quiere potenciar el desarrollo de las diferentes áreas relacionadas con el PLN, mejorar la divulgación de las investigaciones que se llevan a cabo, identificar las futuras directrices de la investigación básica y mostrar las posibilidades reales de aplicación en este campo. Para la realización de este proyecto me he basado en un conjunto de actas disponibles en dicho documento y que procedo a describir:

- **Una cascada de transductores simples para normalizar tweets** (Alegría, Etxeberría, & Labaka, 2013)**.**
  En este caso se presenta un sistema basado en la concatenación de varios transductores o FSTs. Cada uno de los transductores se encarga de completar un hito más o menos simple: ejemplos aprendidos, entidades nombradas, errores básicos, palabras contiguas unidas, onomatopeyas, cambios complejos, cambios en mayúsculas.
  Este proyecto ha sido uno de los pilares en los que he basado mi motor y he tomado muchas referencias de él: onomatopeyas, cambio de mayúscula a minúscula, errores básicos, etc. Además he añadido, por ejemplo, la repetición de caracteres que en este modelo no lo contemplan y es una práctica bastante extendida en las redes sociales. También he introducido herramientas externas con las que ellos no han contado.
- **DLSI en Tweet-norm 2013: Normalización de Tweets en Español** (Mosquera & Moreda, 2013)
  En esta acta presentan TENOR, una herramienta de Normalización Multilingüe. TENOR sigue un proceso de normalización compuesto de dos pasos: En primer lugar se emplea un método de clasificación con el fin de detectar variantes léxicas no-estándar o fuera del vocabulario; En segundo lugar, se sustituyen las palabras seleccionadas en el paso anterior por su forma original normalizada. Utilizan la herramienta externa Freeling para su creación.
  También me he basado en este esquema para la realización de mi motor, en este caso realizo los dos pasos pero utilizando la herramienta de Apicultur en vez de la de Freeling. Uno de los motivos es que, al tratarse de un motor en el que el único idioma es el español, he visto más apropiado utilizar una herramienta especializada en dicho idioma.
- **Análisis lingüístico de expresiones negativas de tweets en español** (Villar Rodriguez, Garcia Serrano, & González Rodríguez, 2013).
  En este artículo se presenta el análisis realizado sobre la colección TASS de tweets en español, para mejorar el análisis automático de su opinión, sobre la base de nuevos recursos léxicos y de procesos de identificación del ámbito y el foco de la negación. Como en esta colección además de las clases de opinión básicas, positiva, negativa, neutral y none, se tienen en cuenta dos nuevas clases muy positiva y muy negativa, se estudian además con detalle los casos en los que intervienen modificadores que afectan (aumentando o invirtiendo) la opinión de un tweet. A lo largo del artículo se presentan ejemplos representativos de cada proceso y se concluye con ejemplos que plantean problemas abiertos aún no incluidos en este trabajo, relacionados con la ironía, la necesidad de contexto y la subjetividad intrínseca de los tweets, que dificultan el análisis de la opinión correcto o al menos consensuado.

Se ha tenido en cuenta la valoración de las expresiones negativas en mi proyecto y se ha presentado el mismo problema sin concluir como es la ironía, la ambigüedad y la subjetividad de los comentarios. Como se comenta a lo largo de esta memoria son problemas con difícil solución, si de por sí es difícil muchas veces entender la ironía de una persona, más difícil será que lo consigan entender este tipo de analizadores.

3 CICLO DE VIDA DEL PROYECTO

Este proyecto se ha desarrollado en base al modelo del ciclo de vida en cascada. Este modelo sigue el siguiente esquema (Berzal):

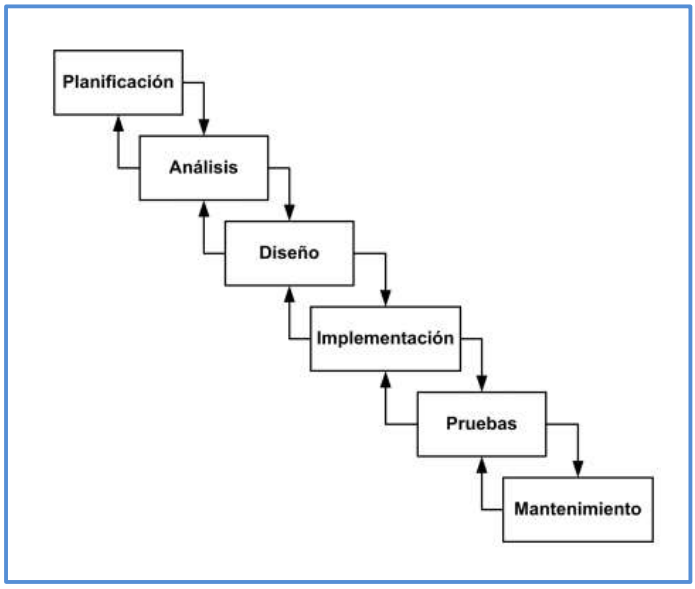


Ilustración 17. Esquema del Ciclo de Vida

En primer lugar se realizó una planificación del sistema, con las fechas estimadas en las que se iba a realizar cada tarea, un presupuesto inicial, y una visión inicial de la viabilidad del proyecto.

En el análisis se realizó un proceso de recopilación de los requisitos tanto funcionales como no funcionales del sistema.

Después se diseñó el sistema centrándose en la estructura de los datos, la arquitectura del software y el detalle procedimental. En esta etapa se hizo una representación gráfica del sistema a partir de los requisitos obtenidos para facilitar la codificación.

Una vez que se realizó el diseño tuvo lugar la implementación.

Cuando se había generado el código se realizaron las pruebas oportunas para verificar el correcto funcionamiento del sistema.

Y una vez que se realizó la aplicación, nos encontramos en la fase de mantenimiento y mejora continua.

En todas estas fases ha habido una labor de documentación continua.

4 METODOLOGÍA

Con el objetivo de crear un sistema de calidad, se ha seguido una metodología durante todo el proyecto.

Una metodología es un conjunto integrado de técnicas y métodos que permite abordar de forma homogénea y abierta cada una de las actividades del ciclo de vida de un proyecto de desarrollo. Es un proceso de software detallado y completo (INTECO, 2009). Una metodología:

- Optimiza el proceso y el producto software
- Define métodos que guían en la planificación y en el desarrollo del software.
- Define qué hacer, cómo y cuándo durante todo el desarrollo y mantenimiento de un proyecto.

Por tanto este proyecto toma como base el estándar IEEE 1074 – 1997 (Bernardos, 2003). Este estándar fue desarrollado por la IEEE para determinar el conjunto de actividades esenciales que deben ser incorporadas en el desarrollo de un producto software.

El IEEE 1074 contempla 17 grupos de actividades y 65 actividades en total. Los grupos de actividades son:

- De Gestión del Proyecto (17 actividades)
  - Iniciación (4 actividades)
  - Planificación (8)
  - Monitoreo y control (5)
- De pre-desarrollo (11)
  - Exploración de conceptos (4)
  - Asignación al Sistema (3)
  - Importación al software (4)
- De desarrollo (10)
  - Requisitos (3)
  - Diseño (4)
  - Implementación (3)
- De post-desarrollo (12)
  - Instalación (3)
  - Operación y soporte (3)
  - Mantenimiento (3)
  - Retiro (3)
- Integrales (15)
  - Evaluación (7)
  - Gestión de configuración (3)
  - Desarrollo de documentación (2)
  - Capacitación (3)

5 ANÁLISIS DEL SISTEMA

En este apartado se va a realizar un análisis del sistema mediante los requisitos de usuario. Estos requisitos se encargarán de dar información acerca de las funcionalidades del sistema y de sus restricciones. Dado que no es el objetivo principal del Trabajo de Fin de Grado se expondrán de manera informal.

Un requisito es una "*condición o capacidad que necesita el usuario para resolver un problema o conseguir un objetivo determinado*". (Landazabal, 2008)

5.1 Requisitos de Usuario

Los requisitos de usuario están divididos en requisitos funcionales y requisitos no funcionales.

Los requisitos funcionales expresan la naturaleza del sistema (como interactúa el sistema con su entorno y cuál va a ser su estado y funcionamiento), los servicios o funciones que proveerá el sistema y describen la interacción entre el sistema y el entorno.

Los requisitos no funcionales son restricciones a los servicios o funciones ofrecidos por el sistema. Describen restricciones que limitan las elecciones para construir una solución.

5.1.1 Requisitos Funcionales

Los requisitos funcionales definidos para este sistema son los siguientes:

Un requisito principal de este proyecto es la obtención de los tweets. Éstos pueden ser obtenidos mediante el API de Twitter o a partir del conjunto de tweets que se nos ha proporcionado para el entrenamiento. Su prioridad es alta.

Una vez que tenemos los tweets es necesario realizar la extracción de los smileys. Se realizará una extracción de los símbolos que imitan sentimientos faciales para poder ser analizados de manera individual. También se deben extraer los #Hashtags, estas palabras vienen precedidas por el símbolo # y también se gestionan de manera individual. Su prioridad es media

Además de extraer algunas de las palabras, es conveniente eliminar otras de ellas, o caracteres que se encuentran dentro de éstas. Se establece como requisito la eliminación de las URLs, todas las palabras que empiecen por "www" o "http://" serán eliminadas del tweet. En cuanto a la eliminación de los caracteres repetidos se eliminarán todos aquellos caracteres repetidos que estén consecutivos, siempre y cuando la palabra no pierda sentido. También se eliminarán los signos de puntuación y demás símbolos que pueden generar error en las herramientas. Su prioridad es media

Con el fin de utilizar herramientas externas será necesario adaptarlas a nuestro trabajo. Para ello se hace necesario la creación de un cliente de la herramienta Apicultur (en concreto

el Lematizador Clásico) y la adaptación del normalizador de la RAE creando un vínculo entre el motor y la herramienta para poder normalizar los tweets. Su prioridad es alta

Será necesario también generar una lista de palabras para compararlas con el tweet y medir el sentimiento de cada palabra y generar un algoritmo que te devuelva el resultado final. Su prioridad es alta

5.1.2 Requisitos No Funcionales

Como requisitos no funcionales se definen en esta memoria la comparación de las diferentes funciones realizadas en la práctica para medir la precisión, el recall y el f-measure de cada una. Su prioridad es alta

La comparación global del sentimiento se realizará a partir del conjunto de entrenamiento proporcionado que se ha comentado anteriormente y de forma general se comparará con la precisión y el recall que han obtenido proyectos anteriores a este.

6 DISEÑO E IMPLEMENTACIÓN DEL SISTEMA

En este apartado se describe el modelo desarrollado durante este trabajo. El objetivo es realizar un analizador de sentimiento de los tweets. Como se ha comentado en el apartado de los objetivos, el programa tiene dos divisiones importantes: la primera es un motor que se encarga de extraer tweets (en tiempo real) de Twitter con su API Streaming y la segunda es el análisis de sentimiento los tweets obtenidos.

Para la realización del motor se nos ha proporcionado un conjunto de alrededor de 600 tweets que venían ya categorizados por el sentimiento. Estos tweets fueron usados para el concurso organizado por el SEPLN del que se ha hablado en el estado del arte y son los que han sido valorados en este motor y nos han servido para comprobar su exactitud.

Puesto que la extracción de tweets se hacía de manera automática a través de una API la definición de la arquitectura se realizará únicamente sobre la segunda parte. De todos modos, en el último apartado de este punto (Ver Apartado 4.6 Tecnologías Usadas) se describirá su funcionamiento.

La arquitectura del análisis del tweet se divide en cuatro módulos:

1. **Tratamiento de las "palabras especiales":** son todas aquellas palabras que no se pueden normalizar porque no tienen sentido de manera autónoma o no pueden ser reconocidas por ningún diccionario pero aportan mucho valor al tweet.
2. **Eliminación de caracteres:** en este módulo se suprimen los signos de puntuación, los caracteres repetidos, etc.
3. **Tratamiento del tweet con herramientas externas:** al tener los tweets "limpios" se normalizan a través de un normalizador de la RAE, se eliminan las palabras que no aportan valor (palabras vacías o nulas) y se usa un lematizador de Apicultur.
4. **Evaluación del sentimiento:** evaluación de cada uno de los tokens, de las palabras extraídas al inicio y valoración del conjunto.

La salida que muestra la aplicación es la valoración del sentimiento del tweet en tres estados:

- Positivo.
- Neutro.
- Negativo.

En el diseño inicial del proyecto se contempló dividir la salida en cinco estados, pero como tal y como se argumenta en el apartado (Ver Apartado 6.2. Ejecucion Final del Proyecto) se propondrá como mejora.

A continuación se muestra un esquema de la arquitectura del proyecto.

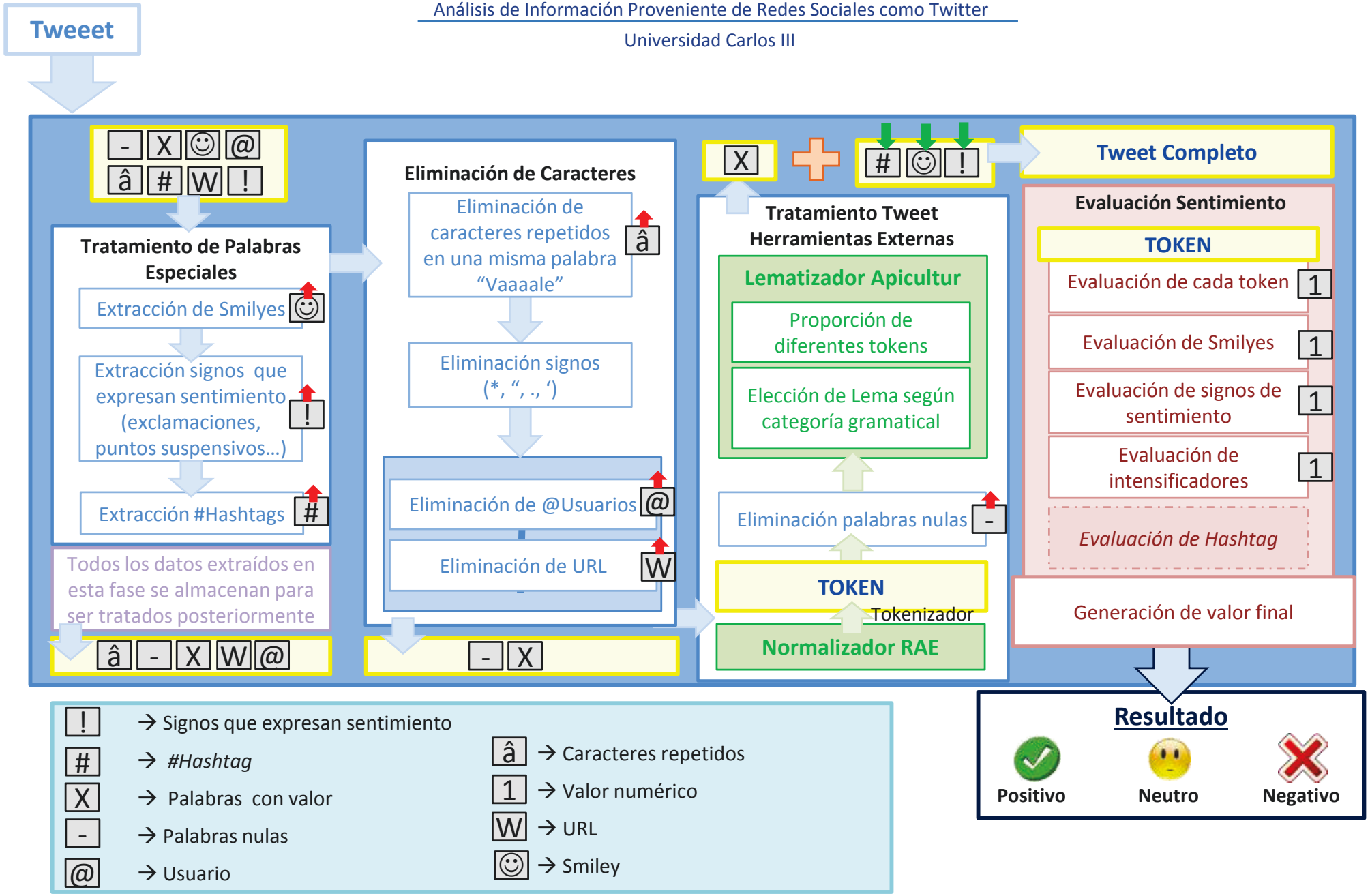


Ilustración 18. . Diseño del Motor de Análisis de Tweets

Una vez visto el diseño completo del motor se va a proceder a la descripción de cada uno de los módulos por separado.

6.1 Tratamiento de palabras especiales

Hemos clasificado como palabras especiales todos aquellos conjuntos de símbolos y letras que no podrían ser reconocidos como palabra en un diccionario pero que aportan mucho valor en un tweet.

Con la llegada de las redes sociales se han popularizado los Smilyes, que son conjuntos de símbolos que unidos forman una *"carita"* y expresan un sentimiento.

Además hay otro tipo de caracteres que pueden aportar mucho valor a una frase sin tener una definición como tal, esto puede ser por ejemplo la onomatopeya de una risa "jajaja", unos puntos suspensivos, un conjunto de exclamaciones, etc.

Por último, un elemento con mucho peso en la redacción de un tweet es el *#Hashtag*. Es un conjunto de caracteres o palabras precedidas de una almohadilla. Sirve como etiquita para señalar el tema sobre el que gira el tweet.

Ilustración 19. Tratamiento de Palabras Especiales

Una vez que se identifican todas estas palabras que aportan mucho valor pero no van a poder ser tratadas por las herramientas existentes en el mercado para el análisis de sentimientos y van a tener que ser tratadas de una manera especial en este TFG, se "sacan" del tweet y se guardan en un array de String para poder ser tratadas posteriormente.

En todo este módulo se trabaja con el tweet completo.

6.2 Eliminación de Caracteres

Una vez que hemos categorizado las palabras especiales pasamos a eliminar los caracteres repetidos dentro de las palabras del tweet.

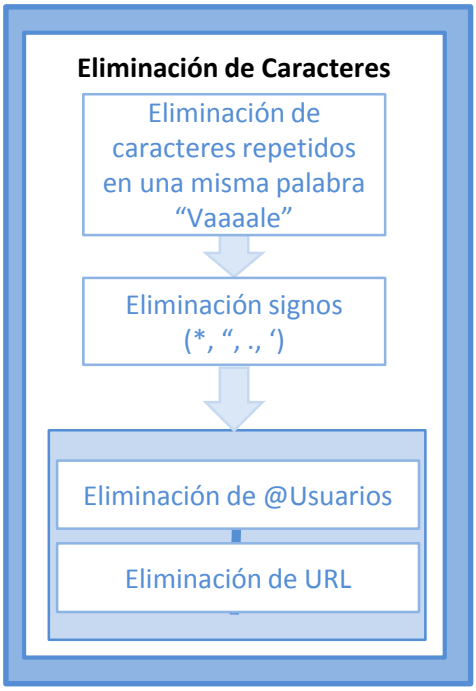


Ilustración 20. Eliminación de Caracteres

Otra práctica muy extendida en la actualidad es la de repetir vocales o ciertas consonantes de algunas palabras, para dar más énfasis.

Para poder tratar estas palabras ha sido necesario eliminar estas repeticiones. Para ello se ha creado un método que detectase si una vocal estaba repetida dos o más veces seguidas y en ese caso se ha eliminado. Ha sido necesario tener en cuenta que hay muchas palabras que tienen dos vocales repetidas seguidas y que la "l" y la "r" a veces son dobles. Se han tratado esas excepciones creando una lista de palabras con esas características y si la palabra era igual a alguna de las de la lista no se eliminaban los caracteres.

Además se ha creado una lista con algunos signos que se han eliminado de la frase. Cierto es, que en algún caso la eliminación de estos símbolos ha cambiado el sentido de la frase y ha hecho que el análisis del sentimiento no fuese del todo acertado, pero si no se eliminaban se producían errores en las herramientas. Otra de las mejoras que se aplicarán a la aplicación será el tratamiento especializado de cada símbolo teniendo en cuenta si la frase cambia de sentido. De todas formas, conseguir que un motor de análisis detecte, por ejemplo, la ironía es una tarea muy complicada, ya de por sí es muy difícil ser detectada por una persona.

Las dos últimas funciones de esta fase son la eliminación de los usuarios (en Twitter todos los usuarios vienen precedidos por el símbolo "@", por lo que en cuanto una palabra empieza con ese símbolo directamente se elimina) y la eliminación de URL (al igual que los usuarios, las URLs vienen precedidas por "*http://*" o "*www*", por tanto también se eliminan). También se ha

pensado en otra mejora que tiene relación con la URL y que se comentará en el apartado correspondiente (Ver Apartado 9. Trabajos Futuros).

En todo este módulo se trabaja con el tweet completo.

6.3 Tratamiento del Tweet con Herramientas Externas

Para continuar con el análisis se han utilizado dos herramientas externas en las que se ha creado un cliente y se ha accedido a ellas para que nos proporcionasen información sobre el tweet. En el apartado (Ver Apartado 4.6. Tecnologías Usadas) se describe como se ha trabajado con cada una de ellas.

Ilustración 21. Tratamiento del Tweet con Herramientas Externas

La línea de trabajo en esta fase ha sido la siguiente:

- En primer lugar el tweet accede al normalizador de la RAE. Éste llega sin usuarios, Hashtags, smileys, etc. ya que habían sido procesados por los dos módulos previos pero entran todas las palabras a la vez, esta herramienta se encargaba de normalizar la frase y de dar coherencia.
- Una vez tratado por el normalizador el tweet se Tokeniza, es decir, todas las palabras se aíslan para ser tratadas por separado. Esto se hace a través del método *StringTokenizer*[2].
- Cada palabra pasa por el método de eliminación de palabras nulas, este método comprueba si la palabra analizada es un artículo, una preposición o una palabra con otra categoría que no aporte valor a la frase, y en el caso de que sea así se elimina. Se comprueba a través de una lista de palabras nulas. En un principio se pensó en pasar

2 La clase *StringTokenizer* nos ayuda a dividir un string en substrings o tokens, en base a otro string (normalmente un carácter) separador entre ellos denominado delimitador.

estas palabras por el Lematizador de Apicultur que indica la categoría a la que pertenece cada palabra, pero este Lematizador tiene el número de llamadas limitado, y la pérdida de capacidad sería bastante grande.

- Cuando ya tenemos las palabras que, en principio, pueden aportar valor a la frase, se pasan por el Lematizador. Antes de lematizarlas se pone toda la palabra en minúscula para poder tratarla mejor de aquí en adelante. La función del Lematizador es la de ofrecerte las palabras de la familia de la palabra que estas introduciendo. Por ejemplo, al introducir la palabra árboles, el lematizador te ofrece los lemas de árbol y arbolar. Además te dice a qué categoría pertenece cada uno de ellos. Lo que se consigue principalmente con este lematizador es saber la categoría de cada palabra para ver el valor que aporta a la frase (en caso de ser adjetivo aporta más que si es un determinante) y simplificarte la palabra a una palabra que puedas tener en tu lista para añadir un valor.
- Por último se vuelve a generar la frase con las palabras que han sido formándose en el lematizador y antes de empezar el módulo siguiente se añadirán los elementos que se han eliminado en la primera fase. El motivo por el que se vuelve a unir la frase completa es simplemente por chequear que el motor está realizando el análisis de forma correcta, podrían pasarse los tokens directamente hasta el final en caso de ser requerido.

Queda como trabajo futuro el uso de otras herramientas que podrían dar valor añadido a este trabajo. Estas herramientas son Freeling (librería gratuita orientada a la prestación de servicios de análisis de lenguaje) y WordNet (base de datos léxica que se encarga de agrupar palabras, proporcionar definiciones cortas y generales, y almacenar las relaciones semánticas entre los conjuntos de sinónimos) entre otras (Ver apartado 9. Trabajos futuros).

6.4 Evaluación del Sentimiento

Llegado a este punto solo queda transformar las palabras en números para poder valorarlos.

Tal y como hemos comentado en el apartado anterior, el tweet llega en forma de frase y es aquí donde se añaden los Smilyes, y los signos de sentimiento. Los *#Hashtags* se valoran de forma neutra hasta que, como también se ha comentado anteriormente, en un trabajo futuro se analicen y se califiquen.

La evaluación de cada token se hace mediante grandes listas de palabras. Estas listas están divididas en palabras muy positivas, palabras muy negativas, palabras positivas y palabras negativas.

Estas listas han sido creadas específicamente para este proyecto. Se han obtenido a partir de la búsqueda de palabras positivas, negativas y sinónimos de cada una de ellas, utilizando distintos diccionarios como fuentes principales[3]. Ha sido un trabajo muy laborioso. Como

[3] Real Academia Española (http://www.rae.es/) y WordReference (http://www.wordreference.com/) principalmente

trabajo futuro (Ver Apartado 9. Trabajos Futuros) se definirá la inclusión en el motor de una herramienta que a partir de una palabra genere los sinónimos y, además de valorar el sentimiento de ésta, vaya añadiendo todos los sinónimos encontrados a cada una de las listas. Cada lista tendrá un valor que será el siguiente:

- Una palabra que este en la lista de palabras muy positivas tendrá el valor 2.
- Una palabra que este en la lista de palabras muy negativas tendrá el valor -2.
- Una palabra que este en la lista de palabras negativas tendrá el valor -1.
- Una palabra que este en la lista de palabras positivas tendrá el valor 1.

Ilustración 22. Evaluación del Sentimiento del Tweet

El funcionamiento para valorar los smileys es distinto. Como hemos comentado en secciones anteriores, los smileys son un conjunto de símbolos que generan una expresión, como por ejemplo: :), ;), =), =(. Están definidos alrededor de sesenta figuras y también tienen un número de categorías limitado, esta vez hay siete categorías.

El objetivo de los smileys es modificar la puntuación obtenida en la frase o en el tweet. Por este motivo es la última acción que se realiza en el análisis. Su funcionamiento es el siguiente: dependiendo de la categoría de smileys se modificará el sentido del tweet de un modo u otro, se describen dos ejemplos aclaratorios: Si el smiley pertenece a una carita sonriente (será la categoría feliz) el tweet aumentará un punto su valoración final, de manera que si la valoración del tweet es 2, ésta será tres (en las categorías similares a ésta se trabaja de la misma forma). En cambio, si el smiley que contiene el tweet es un guiño, la valoración obtenida en el tweet será la contraría, es decir, si el tweet es positivo se convertirá en negativo y viceversa. Se ha decido que con el guiño se cambia el sentido de la frase porque en los tweets estudiados con este símbolo daba a la frase un toque irónico que modificaba el sentido.

Los signos de sentimiento que se han obtenido en la primera fase pertenecen a la onomatopeya de la risa, que esta evaluada con el número 2, los puntos suspensivos que están valorados de forma negativa -2 y los signos de exclamación que también están valorados igual que la risa con 2.

Por último consideramos que los tokens pueden ser intensificadores (palabras como “muy”, “más”, “menos”) para estas palabras también se ha generado una lista de casi cincuenta elementos, y si la palabra coincide con algún intensificador coge el valor de este. La particularidad que tienen los intensificadores es que multiplican la palabra que les precede (en caso de que exista). Como mejora también se propondrá un uso eficiente de este apartado, generando la posibilidad de valorar si el intensificador tiene que multiplicar a la palabra que le precede o que le antecede.

Una vez obtenido el valor de cada uno de los conjuntos se suman y se restan los valores (o se multiplican en el caso de que haya un intensificador) y se obtiene el resultado final.

6.5 Resultado Final

Una vez terminada la evaluación el programa muestra un resultado. Puede ser cualquiera de las tres opciones que se muestran en el cuadro:

Ilustración 23. Resultado del Sentimiento

Como se ha comentado anteriormente el resultado es numérico y la cifra final sale a partir del resultado de las operaciones que se realicen dentro del algoritmo del tweet. Los números proporcionados en la salida son enteros y el resultado se asigna de la siguiente manera:

- Si el valor del tweet es - 2 o menor de - 2 el resultado es negativo.
- Si el valor del tweet oscila entre -1 y 1 el resultado es neutro.
- Si el valor del tweet es 2 o mayor de 2 el resultado es positivo.

6.6 Tecnologías Usadas

En el estado del arte se ha descrito de forma detallada las herramientas que se pueden utilizar para el análisis de texto. En este apartado describiré de forma más técnica como he usado cada una de ellas.

6.6.1 API Streaming de Twitter

La primera herramienta que se utiliza en el proyecto es el API Streaming de Twitter. Esta herramienta es la encargada de proporcionar los tweets que una vez realizado el motor, pasarán a ser analizados.

Para poder utilizar esta API se hace necesario descargar una librería "no oficial" de Twitter llamada *Twitter4j,* disponible para Java. Twitter4j incluye software de JSON para analizar la salida del API.

Se puede elegir los elementos proporcionados en la salida, para este proyecto se implementa la salida con el nombre del usuario y el contenido del tweet. Y también permite que se limite la salida, es decir elegir un conjunto de tweets que tengan algún tipo de relación. En este caso se limita a tweets escritos en español y que contentan en el cuerpo del mensaje una palabra, como puede ser "Nadal", pero esta palabra puede variar a gusto del usuario de la aplicación.

Ilustración 24. Ejemplo API Streaming Twitter

6.6.2 Lematizador Clásico de Apicultur

En el estado del arte comentaba que Apicultur es una plataforma de APIs y para este proyecto se ha usado un API en concreto que es el Lematizador Clásico.

Su función es la siguiente: se introduce una palabra (String) en el Lematizador y éste devuelve todos los lemas posibles de la palabra introducida junto a su categoría gramatical, según la clasificación de Molino de Ideas y según el listado clásico de categorías gramaticales. Apicultur te ofrece una lista donde se encuentran los códigos de estas categorías gramaticales. El formato de la salida es JSON.

Para utilizar este lematizador es necesario registrarte en la página y añadir la API a tus aplicaciones. Una vez añadido te dan unas claves para poder utilizarla. Tienes además que insertar la URL y el path correspondiente al lematizador. Este caso son los siguientes: "*http://store.apicultur.com"* (URL) y *"/api/lematiza-clasico/1.0.0"* (Path).

A continuación se muestra un ejemplo:

Ejemplo de Entrada

http://store.apicultur.com/api/lematiza-clasico/1.0.0/leyendo

Ejemplo de Salida

{"palabra":"leyendo","lemas":[{"lema":"leer","categoria":"25","categoriaSimple":"5"}]}

En el ejemplo anterior los parámetros de entrada y la salida que se produce serían los siguientes:

Parámetros de Entrada:

Parámetro 1

Nombre	palabra
Tipo	String
Descripción	Palabra que se quiere lematizar (de la que se quiere saber el lema)

Ilustración 25. Parámetros de entrada de la herramienta Apicultur

Salida

La salida es el objeto JSON que contiene:

Nombre	"palabra"
Tipo	String
Descripción	Palabra que queremos lematizar

Nombre	"lemas"
Tipo	Array
Descripción	Lista que contiene los lemas posibles de la palabra

Nombre	"lema"
Tipo	String
Descripción	Lema propuesto.

Nombre	"categoria"
Tipo	Integer
Descripción	La categoría gramatical del lema propuesto (según la clasificación de Molino de Ideas)

Nombre	"categoriaSimple"
Tipo	Integer
Descripción	La categoría gramatical de la palabra que queremos lematizar (según la clasificación clásica).

Ilustración 26. Salida Lematizador Apicultur

6.6.3 Normalizador RAE

La Real Academia de la lengua Española tiene un grupo de investigación relacionado con la normalización de textos. En concreto han puesto recientemente de forma pública (aunque aún solo con acceso restringido para pruebas) el normalizador de tweets. Al igual que con la API anterior hay para utilizar este servicio es necesario pasar un path con la frase que se quiere normalizar. El path es el siguiente: “http://193.145.222.16/tweet-norm?text=”

En este caso en vez de pasar una sola palabra se envía una frase completa para poder normalizar en función del lugar que ocupen cada una de las palabras.

Se muestra un ejemplo a continuación:

Ejemplo de entrada

“paula viene A madrid mañana En avion”

Ejemplo de salida

“Paula viene a Madrid mañana en avión”

La salida también es un objeto JSON.

En la página siguiente encontramos una imagen con el esquema del proyecto, esta vez se ha insertado un ejemplo para tener una visión del funcionamiento de la aplicación.

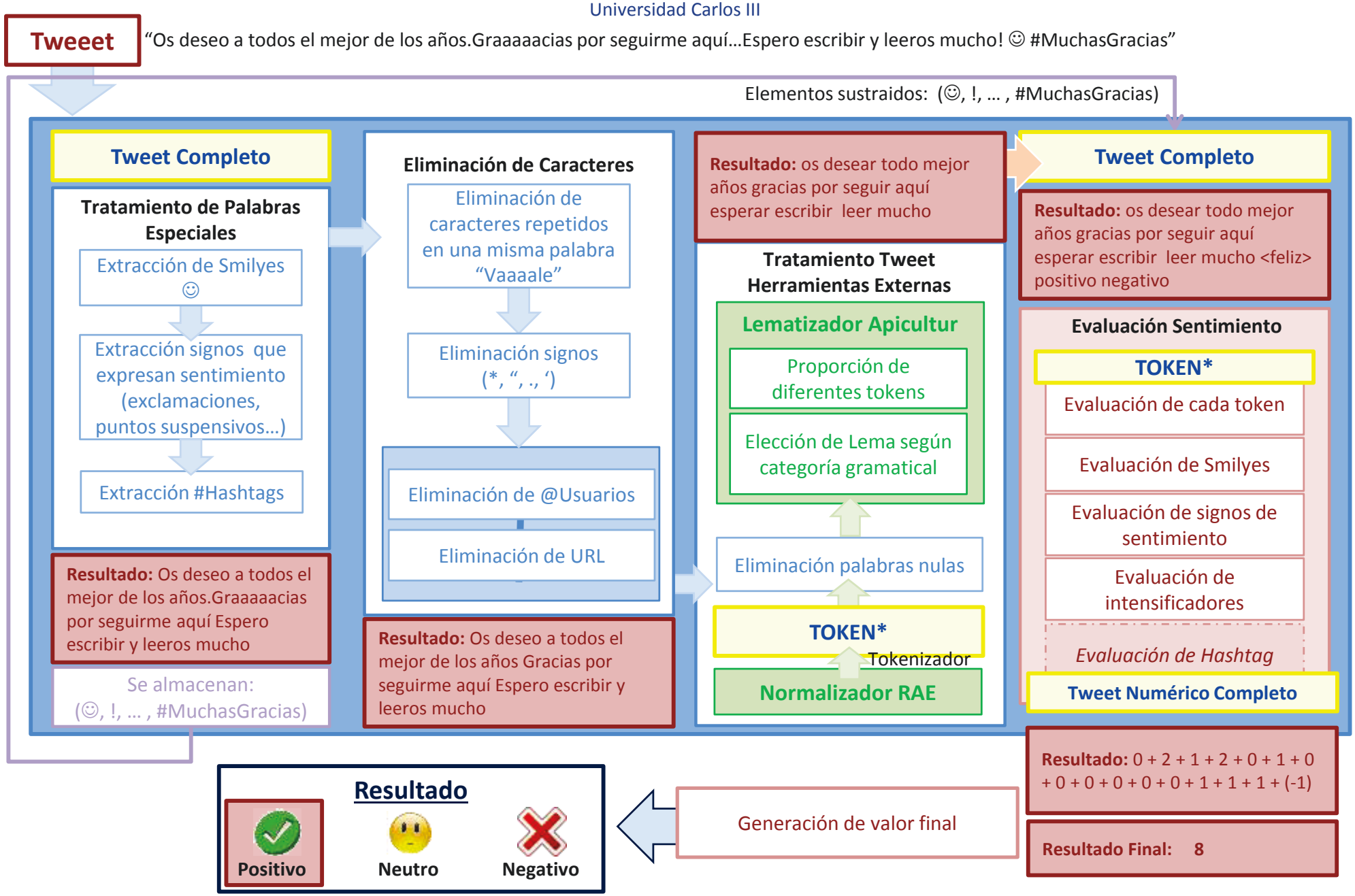


Ilustración 27. Ejemplo Análisis de un Tweet

7 PRUEBAS

Para comprobar el funcionamiento del motor se han realizado una serie de pruebas tanto unitarias como generales. Para cada prueba se han ido cambiando tanto los tweets como los parámetros adaptándose a las entradas y las salidas.

7.1 Pruebas Unitarias

Se han realizado pruebas unitarias de las herramientas externas incluidas en el proyecto para verificar su funcionamiento.

7.1.1 Pruebas Lematizador Apicultur

A continuación se muestran una serie de pruebas realizadas con la herramienta de Apicultur. En esta primera prueba el funcionamiento es correcto, muestra distintas categorías y lemas de cada una de las palabras del tweet.

```
El tweet que se va a lematizar es: Cambio de jefe de prensa en la Zarzuela
{"categoria":"5","categoriaSimple":"5","lema":"cambiar"}
{"categoria":"4","categoriaSimple":"4","lema":"cambio"}
{"categoria":"4","categoriaSimple":"4","lema":"de"}
{"categoria":"9","categoriaSimple":"9","lema":"de"}
{"categoria":"23","categoriaSimple":"4","lema":"jefe"}
{"categoria":"4","categoriaSimple":"4","lema":"de"}
{"categoria":"9","categoriaSimple":"9","lema":"de"}
{"categoria":"5","categoriaSimple":"5","lema":"prensar"}
{"categoria":"27","categoriaSimple":"5","lema":"prensar"}
{"categoria":"4","categoriaSimple":"4","lema":"prensa"}
{"categoria":"9","categoriaSimple":"9","lema":"en"}
{"categoria":"4","categoriaSimple":"4","lema":"la"}
{"categoria":"6","categoriaSimple":"6","lema":"lo"}
{"categoria":"16","categoriaSimple":"16","lema":"el"}
{"categoria":"4","categoriaSimple":"4","lema":"zarzuela"}
```

Ilustración 28. Prueba I, Lematizador Apicultur

Antes de que el tweet pase por esta herramienta es necesario eliminar todos los signos de puntación porque si no se producen errores. Además una vez que se produce el error en una palabra ya no “remonta” la herramienta hasta que pasa al siguiente tweet. Se puede ver un ejemplo a continuación:

```
El tweet que se va a lematizar es: Un abrazo a tod@s.. Y gracias por la buena energía.
GET http://store.apicultur.com/api/lematiza-clasico/1.0.0/Un returned a response status of 200 OK
{"categoria":"16","categoriaSimple":"16","lema":"un"}
GET http://store.apicultur.com/api/lematiza-clasico/1.0.0/abrazo returned a response status of 200 OK
{"categoria":"5","categoriaSimple":"5","lema":"abrazar"}
{"categoria":"4","categoriaSimple":"4","lema":"abrazo"}
GET http://store.apicultur.com/api/lematiza-clasico/1.0.0/a returned a response status of 200 OK
{"categoria":"4","categoriaSimple":"4","lema":"a"}
{"categoria":"9","categoriaSimple":"9","lema":"a"}
GET http://store.apicultur.com/api/lematiza-clasico/1.0.0/tod@s.. returned a response status of 500 Internal Server Error
GET http://store.apicultur.com/api/lematiza-clasico/1.0.0/Y returned a response status of 500 Internal Server Error
GET http://store.apicultur.com/api/lematiza-clasico/1.0.0/gracias returned a response status of 500 Internal Server Error
GET http://store.apicultur.com/api/lematiza-clasico/1.0.0/por returned a response status of 500 Internal Server Error
GET http://store.apicultur.com/api/lematiza-clasico/1.0.0/la returned a response status of 500 Internal Server Error
GET http://store.apicultur.com/api/lematiza-clasico/1.0.0/buena returned a response status of 500 Internal Server Error
GET http://store.apicultur.com/api/lematiza-clasico/1.0.0/energ%C3%ADa. returned a response status of 500 Internal Server Error
```

Ilustración 29. Prueba II, Lematizador Apicultur

7.1.2 Pruebas Normalizador de la RAE

A continuación se muestran las pruebas realizadas con el normalizador de la RAE.

En la primera prueba se puede observar cómo el lematizador se encarga de poner la primera letra en mayúscula de los nombres propios y también añade tildes a las palabras que las deberían llevar y que no se incluyeron a la hora de escribir el tweet (un error muy habitual en este tipo de redes sociales).

```
El tweet que se va a lematizar es: hay en madrid un camion aparcado en doble fila por la zona de San Blas
hay en Madrid un camión aparcado en doble fila por la zona de San Blas
```

Ilustración 30. Prueba III, Analizador RAE

En la segunda prueba se puede comprobar cómo el normalizador de RAE también tiene la funcionalidad de modificar las palabras abreviadas y dar una solución no siempre acertada, de las palabras que están mal escritas.

```
El tweet que se va a lematizar es: no he aparcdo bien el coxe xq no me an dejado
No he aparecido bien el coche porque no me han dejado
```

Ilustración 31. Prueba IV, Normalizador RAE

7.2 Pruebas generales

A continuación se van a realizar dos partes distintas, en la primera se van a mostrar un conjunto de tweets de los que se van a obtener los siguientes valores:

- Número de tweets que vamos a analizar (cogeremos un pequeño conjunto de 10 que nos permita ver los resultados de forma óptima).
- Información por cada tweet:
  - Usuario y contenido del tweet.
  - Valoración del tweet generada por nuestro motor.
  - Valoración del tweet por parte de TASS
  - Información de si hemos acertado o no.

```
CarmendelRiego:@marodriguezb Gracias MAR
 Valor mi analizador:P
 Valor TASS:P
Acierto
-------------------------------------
mgilguerrero:Off pensando en el regalito Sinde, la que se va de la SGAE cuando se van sus corruptos. Intento no sacar conclusiones (lo intento)
 Valor mi analizador:N
 Valor TASS:N+
Acierto
-------------------------------------
paurubio:Conozco a alguien q es adicto al drama! Ja ja ja te suena d algo!
 Valor mi analizador:P
 Valor TASS:P+
Acierto
-------------------------------------
Ignacos:Hoy asisitiré en Madrid a un seminario sobre la Estrategia Española de Seguridad organizado por FAES.
 Valor mi analizador:NEU
 Valor TASS:NONE
Acierto
-------------------------------------
nacho_uriarte:Buen día todos! Lo primero mandar un abrazo grande a Miguel y a su familia @libertadmontes Hoy podría ser un día para la grandeza humana.
 Valor mi analizador:P
 Valor TASS:P+
Acierto
-------------------------------------
JuanraLucas:Desde el escaño. Todo listo para empezar #endiascomohoy en el Congreso http://t.co/Mu2yIgCb
 Valor mi analizador:NEU
 Valor TASS:P+
Fallo
-------------------------------------
Carlos_Latre:"@adri_22_22: #programascambiados es TT gracias a @Carlos_Latre" GRACIAS POR EL BUEN RATO AMIGOS!!;)
 Valor mi analizador:P
 Valor TASS:P+
Acierto
-------------------------------------
mariviromero:Vamos a por el viernes (@ Ayuntamiento de Málaga) [pic]: http://t.co/lzDVsoCu
 Valor mi analizador:NEU
 Valor TASS:NONE
Acierto
-------------------------------------
anabelenroy_tve:Accidente en BUS-VAO A-6 km. 12. Motorista de 30 años herido menos grave. @SAMUR_PC traslada a Hospital Doce de Octubre vía @EmergenciasMad
 Valor mi analizador:NEU
 Valor TASS:N
Fallo
-------------------------------------
El número de tweets es que se ha analizado es: 10tweets
He acertado el sentimiento de  7 tweets
```

Ilustración 32. Prueba V, Prueba General

En la siguiente imagen se va a mostrar un conjunto de valores obtenidos tras realizar varias pruebas con el motor. El método de actuación ha sido el siguiente: se han generado distintos grupos de tweets con un tamaño aleatorio de aproximadamente 60 tweets y se han ido analizando a lo largo de los últimos días del proyecto. El motivo de que se haya realizado el análisis en grupos reducidos es evitar la saturación de las herramientas externas utilizadas pero el análisis de todos los tweets se ha hecho con las mismas condiciones. Han sido necesarias muchas horas para la validación de los tweets debido a la cantidad de información gestionada. Se ha dividido la solución en tres tablas en las que vamos a poder observar los aciertos y los fallos que se han producido en el análisis de los tweets y la precisión, el recall y el F-measure.

En esta primera tabla medimos la precisión y mostramos la siguiente información:

- ID del conjunto.
- Número total de tweets analizados.
- Medición de la precisión con los siguientes datos:
  - Tweets marcados como positivos (sombreado en azul)
  - Tweets marcados como negativos (sombreado en azul)
  - Tweets marcados como neutros (sombreado en azul)
    Todos ellos tienen la siguiente información:
    - Número de Aciertos (TP)
    - Número de Fallos (FP)
- Por último valor de la precisión.

		Precisión						
Conjunto	**Nº Tweets**	**Positivo**		**Negativo**		**Neutro**		**Precisión**
		Aciertos	**Fallos**	**Aciertos**	**Fallos**	**Aciertos**	**Fallos**	
GRUPO001	67	23		32		12		0,6119403
		18	5	18	14	5	7	
GRUPO002	69	28		24		17		0,56521739
		19	9	14	10	6	11	
GRUPO003	64	14		42		8		0,859375
		12	2	36	6	7	1	
GRUPO004	58	17		30		11		0,55172414
		11	6	15	15	6	5	
GRUPO005	71	43		6		22		0,56338028
		25	18	5	1	10	12	
GRUPO006	64	20		27		17		0,609375
		14	6	10	17	15	2	
GRUPO007	66	13		12		41		0,60606061
		9	4	8	4	23	18	
GRUPO008	58	6		39		13		0,46551724
		6	0	11	28	10	3	
GRUPO009	58	10		13		35		0,62068966
		6	4	3	10	27	8	
GRUPO010	63	33		25		5		0,65079365
		27	6	10	15	4	1	
TOTAL								**0,7015914**

Ilustración 33. Prueba VI, Precisión

En la segunda tabla medimos el recall y mostramos la siguiente información:

- ID del conjunto.
- Número total de tweets analizados.
- Medición del recallcon los siguientes datos:
  - Total de tweets positivos (sombreado en azul)
  - Total de tweets negativos (sombreado en azul)
  - Total de tweets neutros (sombreado en azul)
    Todos ellos tienen la siguiente información:
    - Número de Aciertos (TP)
    - Número de Fallos (FN)
- Por último valor del recall.

		Recall						
Conjunto	**Nº Tweets**	**Positivo**		**Negativo**		**Neutro**		**Recall**
		Aciertos	**Fallos**	**Aciertos**	**Fallos**	**Aciertos**	**Fallos**	
GRUPO001	67	34		21		12		0,74626866
		26	8	16	5	8	4	
GRUPO002	69	27		32		10		0,65217391
		22	5	17	15	6	4	
GRUPO003	64	14		38		12		0,828125
		10	4	35	3	8	4	
GRUPO004	58	16		33		9		0,86206897
		13	3	32	1	5	4	
GRUPO005	71	30		28		13		0,77464789
		22	8	21	7	12	1	
GRUPO006	64	18		24		22		0,78125
		16	2	17	7	17	5	
GRUPO007	66	15		24		27		0,87878788
		12	3	19	5	27	0	
GRUPO008	58	20		23		15		0,74137931
		12	8	19	4	12	3	
GRUPO009	58	10		16		32		0,77586207
		7	3	8	8	30	2	
GRUPO010	63	35		21		7		0,84126984
		22	13	27	-6	4	3	
TOTAL								**0,78818335**

Ilustración 34. Prueba VII, recall

En la última tabla se muestra la relación entre la precisión y el recall calculada con el F-measure.

Conjunto	Nº Tweets	Precisión	Recall	F-measure
GRUPO001	67	0,6119403	0,74626866	0,67246187
GRUPO002	69	0,56521739	0,65217391	0,60559006
GRUPO003	64	0,859375	0,828125	0,84346065
GRUPO004	58	0,55172414	0,86206897	0,67283431
GRUPO005	71	0,56338028	0,77464789	0,65233506
GRUPO006	64	0,609375	0,78125	0,68469101
GRUPO007	66	0,60606061	0,87878788	0,71737786
GRUPO008	58	0,46551724	0,74137931	0,57192118
GRUPO009	58	0,62068966	0,77586207	0,68965517
GRUPO010	63	0,65079365	0,84126984	0,73387369
TOTAL		**0,7015914**	**0,78818335**	**0,68442009**

Ilustración 35. Prueba VIII, F-measure

Como se puede ver en la última tabla los valores obtenidos en el recall son mejores que los obtenidos en la precisión. Esto se debe a que se ha intentado que todos los tweets que fuesen tanto positivos, negativos como neutros estuvieran bien categorizados y ha supuesto la introducción de algo de ruido en el resultado final.

El aumento de las listas de palabras y el uso de más herramientas externas podría favorecer el aumento de estos valores. De todas formas, como se ha comentado a lo largo de todo el proyecto, hay una variante con la que es muy difícil competir y es la ironía y los juegos de palabras, este análisis de hace en base a las palabras obtenidas pero no es capaz de captar esos matices.

8 PLANIFICACIÓN

Este apartado está dividido en dos secciones: en la primera se muestra una planificación que se hizo inicialmente de manera orientativa y en la segunda se definirá la ejecución del proyecto que se ha realizado en realidad.

Las variaciones se deben a ciertos problemas técnicos que han surgido durante la elaboración del proyecto y a nuevas ideas que han ido surgiendo y se han ido adaptando a lo largo de este.

Para ambas planificaciones se ha utilizado el diagrama de Gantt (Vallejo, 2013). Ésta es una herramienta gráfica cuyo objetivo es exponer el tiempo de dedicación previsto para diferentes tareas o actividades a lo largo de un tiempo total determinado. Es importante tener en cuenta que el diagrama de Gantt no indica las relaciones existentes entre actividades, solo el tiempo invertido en cada una de ellas.

Este diagrama ha sido creado con la herramienta MS Project[4]. Es un software de administración de proyectos diseñado, desarrollado y comercializado por Microsoft para asistir a administradores de proyectos en el desarrollo de planes, asignación de recursos a tareas, dar seguimiento al progreso, administrar presupuesto y analizar cargas de trabajo.

En los gráficos obtenidos con el Project no se puede obtener una visión real del tiempo invertido en cada fase, ya que por motivos laborales no puedo invertir el mismo número de horas cada día. Por ello también se realiza un gráfico con las horas invertidas en casa fase.

En la elaboración de este proyecto han trabajado Ana Iglesias Maqueda (Tutora del Proyecto) y Soledad Martín Morales (Desarrolladora del Proyecto).

La planificación se ha realizado en base a las fases necesarias para el ciclo de vida de un proyecto (Berzal, El ciclo de vida de un sistema de información). Estas fases son:

5. **Planificación:** Antes de iniciar un proyecto es necesario realizar una serie de tareas previas que influirán decisivamente en la finalización con éxito del proyecto. Estas tareas se conocen popularmente como el *fuzzy front-end* del proyecto al no estar sujetas a plazos. Las tareas iniciales que se realizarán esta fase inicial del proyecto incluyen actividades tales como la determinación del ámbito del proyecto, la realización de un estudio de viabilidad, una estimación del coste del proyecto, su planificación temporal y la asignación de recursos a las distintas etapas del proyecto.
6. **Análisis:** Lo primero que se debe hacer cuando quieres construir un sistema de información es averiguar qué es exactamente lo que tiene que hacer el sistema. La etapa de análisis en el ciclo de vida del software corresponde al proceso mediante el cual se intenta descubrir qué es lo que realmente se necesita y se llega a una comprensión adecuada de las características que el

[4] Microsoft Project Standar 2013. http://www.microsoftstore.com

sistema debe poseer. En esta fase se deben analizar los requerimientos del programa y se debe analizar la situación actual del ámbito del proyecto.

7. **Diseño:** Mientras que los modelos utilizados en la etapa de análisis representan los requisitos del usuario desde distintos puntos de vista (el qué), los modelos que se utilizan en la fase de diseño representan las características del sistema que nos permitirán implementarlo de forma efectiva (el cómo).
8. **Implementación:** Una vez que sabemos qué funciones debe desempeñar nuestro sistema de información (análisis) y hemos decidido cómo vamos a organizar sus distintos componentes (diseño), es el momento de pasar a la etapa de implementación.
9. **Pruebas:** la etapa de pruebas tiene como objetivo detectar los errores que se hayan podido cometer en las etapas anteriores del proyecto.

Además de estas fases se realiza una tarea de documentación a lo largo de todo el proyecto.

En cada una de las fases para las dos planificaciones se detallará el número de horas invertidas y las personas que han participado.

8.1 PLANIFICACIÓN INICIAL

En esta planificación inicial se indican las fases necesarias para el ciclo de vida del proyecto. La planificación inicial estaba definida con una duración de 300h que es la duración estimada por la Universidad Carlos III para la realización del Proyecto de Fin de Grado.

1. **Planificación.**
   - 1.1. Determinación del ámbito del proyecto
     - Tiempo estimado: 4 horas.
     - Participantes: Ana Iglesias Maqueda y Soledad Martín Morales.
   - 1.2. Realización de un estudio de viabilidad
     - Tiempo estimado: 5 horas.
     - Participantes: Ana Iglesias Maqueda y Soledad Martín Morales.
   - 1.3. Estimación del coste del proyecto
     - Tiempo estimado: 2 horas.
     - Participantes: Soledad Martín Morales.
   - 1.4. Planificación temporal
     - Tiempo estimado: 2 horas.
     - Participantes: Soledad Martín Morales.
   - 1.5. Asignación de recursos
     - Tiempo estimado: 1 horas.
     - Participantes: Soledad Martín Morales.

El total de horas estimadas para la planificación es: 13 horas.

2. **Análisis**
   - 2.1. Análisis del estado del arte

- Tiempo estimado: 70 horas.
- Participantes: Soledad Martín Morales.

2.2. Análisis de los requerimientos del programa:
- Tiempo estimado: 15 horas.
- Participantes: Ana Iglesias Maqueda y Soledad Martín Morales.

El total de horas estimadas para el análisis es: 85 horas.

3. Diseño

3.1. Definición de la arquitectura:
- Tiempo estimado: 4 horas.
- Participantes: Soledad Martín Morales.

3.2. Diseño de la interoperabilidad de las herramientas
- Tiempo estimado: 11 horas.
- Participantes: Soledad Martín Morales.

El total de horas estimadas para el diseño es: 15 horas.

4. Implementación

4.1. Desarrollo de los clientes para las APIs utilizadas:
- Tiempo estimado: 35 horas.
- Participantes: Soledad Martín Morales.

4.2. Desarrollo de los métodos de análisis:
- Tiempo estimado: 30 horas.
- Participantes: Soledad Martín Morales.

4.3. Desarrollo de los métodos de clasificación:
- Tiempo estimado: 15 horas.
- Participantes: Soledad Martín Morales.

4.4. Desarrollo de los algoritmos de validación:
- Tiempo estimado: 10 horas.
- Participantes: Soledad Martín Morales.

El total de horas estimadas para el diseño es: 90 horas.

5. Pruebas

5.1. Pruebas unitarias de cada herramienta instalada.
- Tiempo estimado: 20 horas.
- Participantes: Soledad Martín Morales.

5.2. Pruebas globales
- Tiempo estimado: 20 horas.
- Participantes: Soledad Martín Morales.

El total de horas estimadas para el diseño es: 40 horas.

6. Documentación

6.1. Documentación del proyecto:
- Tiempo estimado: 50 horas.

- Participantes: Soledad Martín Morales.

6.2. Presentación:

- Tiempo estimado: 7 horas.
- Participantes: Soledad Martín Morales.

El total de horas estimadas para la documentación es de 57 horas.

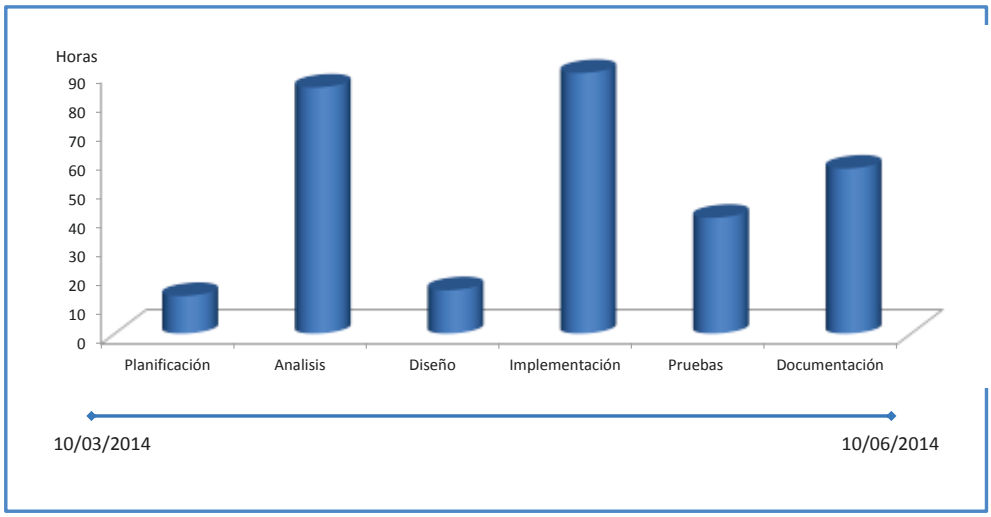
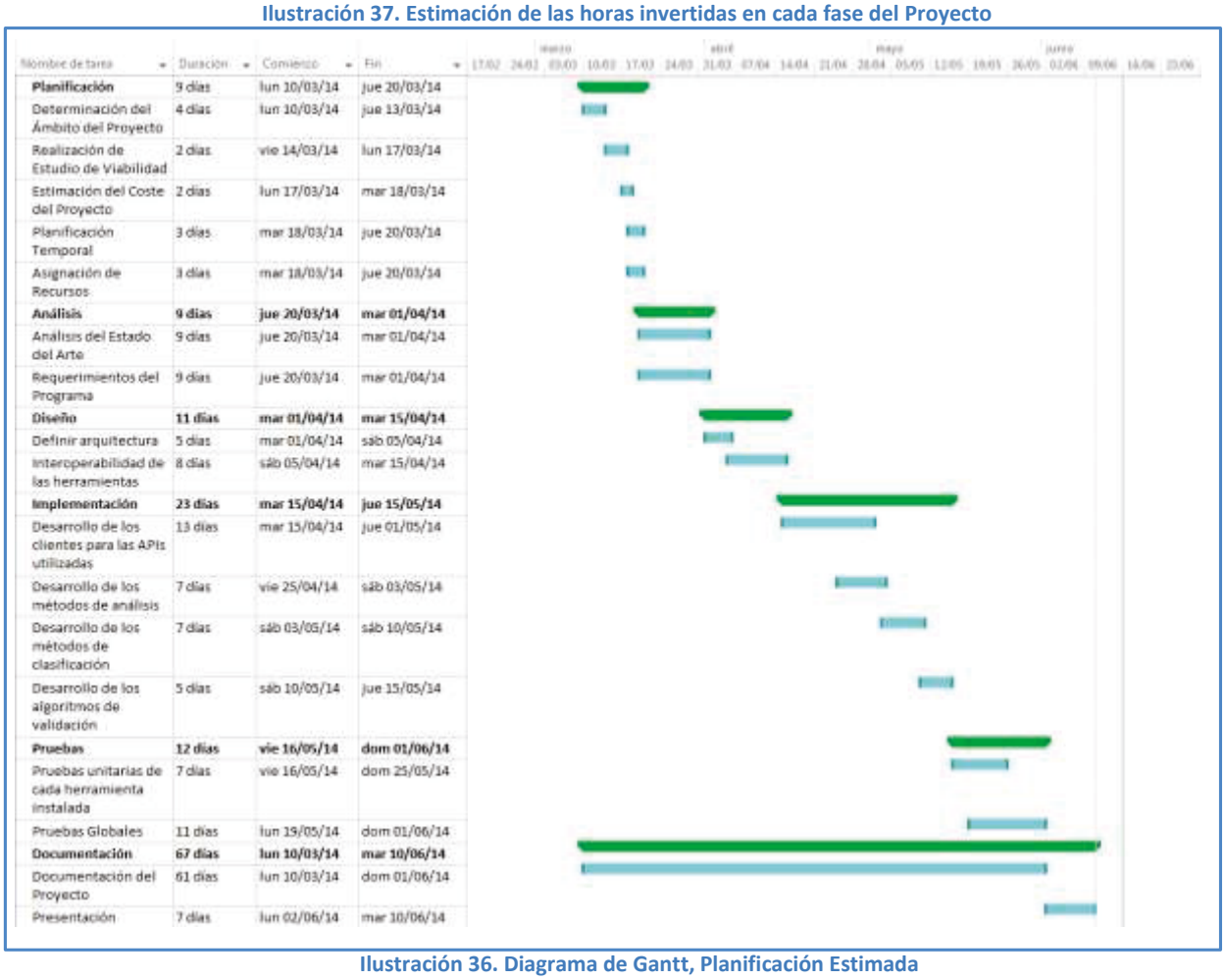


Ilustración 37. Estimación de las horas invertidas en cada fase del Proyecto



Nombre de tarea	Duración	Comienzo	Fin
Planificación	9 días	lun 10/03/14	jue 20/03/14
Determinación del Ámbito del Proyecto	4 días	lun 10/03/14	jue 13/03/14
Realización de Estudio de Viabilidad	2 días	vie 14/03/14	lun 17/03/14
Estimación del Coste del Proyecto	2 días	lun 17/03/14	mar 18/03/14
Planificación Temporal	3 días	mar 18/03/14	jue 20/03/14
Asignación de Recursos	3 días	mar 18/03/14	jue 20/03/14
Análisis	**9 días**	**jue 20/03/14**	**mar 01/04/14**
Análisis del Estado del Arte	9 días	jue 20/03/14	mar 01/04/14
Requerimientos del Programa	9 días	jue 20/03/14	mar 01/04/14
Diseño	**11 días**	**mar 01/04/14**	**mar 15/04/14**
Definir arquitectura	5 días	mar 01/04/14	sáb 05/04/14
Interoperabilidad de las herramientas	8 días	sáb 05/04/14	mar 15/04/14
Implementación	**23 días**	**mar 15/04/14**	**jue 15/05/14**
Desarrollo de los clientes para las APIs utilizadas	13 días	mar 15/04/14	jue 01/05/14
Desarrollo de los métodos de análisis	7 días	vie 25/04/14	sáb 03/05/14
Desarrollo de los métodos de clasificación	7 días	sáb 03/05/14	sáb 10/05/14
Desarrollo de los algoritmos de validación	5 días	sáb 10/05/14	jue 15/05/14
Pruebas	**12 días**	**vie 16/05/14**	**dom 01/06/14**
Pruebas unitarias de cada herramienta instalada	7 días	vie 16/05/14	dom 25/05/14
Pruebas Globales	11 días	lun 19/05/14	dom 01/06/14
Documentación	**67 días**	**lun 10/03/14**	**mar 10/06/14**
Documentación del Proyecto	61 días	lun 10/03/14	dom 01/06/14
Presentación	7 días	lun 02/06/14	mar 10/06/14

Ilustración 36. Diagrama de Gantt, Planificación Estimada

8.2 EJECUCIÓN FINAL DEL PROYECTO

En esta sección se indican tanto las fases necesarias para el ciclo de vida del proyecto como la descripción detallada de cada una de ellas. Además se comentará por qué se ha producido el desfase de las horas iniciales.

1. **Planificación.**
   - 1.1. Determinación del ámbito del proyecto: En varias reuniones mantenidas con la tutora del proyecto en las que me explicó una gran diversidad de acciones que se podían realizar con el tema elegido determinamos que el ámbito del proyecto se iba a centrar en analizar el sentimiento de los tweets.
     - Tiempo invertido: 4 horas.
     - Participantes: Ana Iglesias Maqueda y Soledad Martín Morales.
   - 1.2. Realización de un estudio de viabilidad. A raíz de esas reuniones se buscó información para estudiar la viabilidad del proyecto.
     - Tiempo invertido: 5 horas.
     - Participantes: Ana Iglesias Maqueda y Soledad Martín Morales.
   - 1.3. Estimación del coste del proyecto. Se hizo una estimación inicial teniendo en cuenta unas herramientas iniciales que han tenido que ser modificadas y con los recursos que en este caso éramos nosotras dos.
     - Tiempo invertido: 2 horas.
     - Participantes: Soledad Martín Morales.
   - 1.4. Planificación temporal. De manera orientativa también se realizó la planificación que se ha descrito anteriormente.
     - Tiempo invertido: 2 horas.
     - Participantes: Soledad Martín Morales.
   - 1.5. Asignación de recursos. Se asignaron los recursos a todas las fases del proyecto.
     - Tiempo invertido: 1 horas.
     - Participantes: Soledad Martín Morales.

El total de horas estimadas que duró la planificación no supuso ningún cambio en relación con lo que estaba estimado. Esta duración fue de 14 horas.

2. **Análisis**
   - 2.1. Análisis del estado del arte. Antes de crear el motor era necesario tener una visión actual de las herramientas que existían ya en el mercado para la realización de temas parecidos a los del proyecto. Ésta fue una gran tarea de investigación y documentación que se ha desarrollado a lo largo de todo el proyecto.
     - Tiempo invertido: 78 horas.
     - Participantes: Soledad Martín Morales.
   - 2.2. Análisis de los requerimientos del programa: para la realización del proyecto era necesario utilizar una serie de herramientas y cada una de ellas necesitaba ser analizada.
     - Tiempo invertido: 10 horas.

- Participantes: Ana Iglesias Maqueda y Soledad Martín Morales.

El total de horas invertidas en el análisis es: 88 horas.

3. Diseño

3.3. Definición de la arquitectura: En esta fase se diseña la arquitectura del proyecto.
- Tiempo invertido: 7 horas.
- Participantes: Soledad Martín Morales.

3.4. Diseño de la interoperabilidad de las herramientas: Como se van a utilizar distintas herramientas en el proyecto se diseña la manera en la que van a interactuar.
- Tiempo invertido: 10 horas.
- Participantes: Soledad Martín Morales.

El total de horas estimadas para el diseño es: 17 horas.

4. Implementación

4.5. Desarrollo de los clientes para las APIs utilizadas: Este es el punto más crítico de todo el proyecto. En el diseño y el análisis se propone realizar el proyecto en Linux y con Eclipse (programado en Java). Pero muchas de las APIs que se iban a utilizar no funcionan, o no tienen la funcionalidad en español como venía en su descripción por lo que es necesario cambiar de herramientas y volver a estudiar el funcionamiento de cada una de ellas. Por motivos de operatividad se cambia de sistema operativo y esto hace que se retrase el proyecto un largo periodo de tiempo.
- Tiempo invertido: 60 horas.
- Participantes: Soledad Martín Morales.

4.6. Desarrollo de los métodos de análisis: Los métodos de análisis se realizan a la par que el desarrollo de los clientes por lo que también sufre variación.
- Tiempo invertido: 35 horas.
- Participantes: Soledad Martín Morales.

4.7. Desarrollo de los métodos de clasificación: Se limitan los métodos de clasificación a partir de los que había diseñados. Siguen realizando la misma función pero no se obtiene tanto detalle.
- Tiempo invertido: 10 horas.
- Participantes: Soledad Martín Morales.

4.8. Desarrollo de los algoritmos de validación: Los algoritmos de validación siguen la planificación incial.
- Tiempo invertido: 10 horas.
- Participantes: Soledad Martín Morales.

El total de horas estimadas para el diseño es: 115 horas.

5. Pruebas

5.3. Pruebas unitarias de cada herramienta instalada: Únicamente se realizan pruebas con las herramientas válidas por lo que tampoco difiere en exceso de la planificación inicial.
- Tiempo invertido: 16 horas.
- Participantes: Soledad Martín Morales.

5.4. Pruebas globales: Por miedo a que los servidores dejen de funcionar en momentos críticos se realizan más pruebas de las que se habían planificado en un principio.
- Tiempo invertido: 21 horas.
- Participantes: Soledad Martín Morales.

El total de horas estimadas para el diseño es: 37 horas.

6. Documentación

6.3. Documentación del proyecto: La documentación del proyecto se realiza a lo largo de todo el proyecto junto con el análisis del estado del arte, están fuertemente conectados el uno con el otro.
- Tiempo invertido: 60 horas.
- Participantes: Soledad Martín Morales.

6.4. Presentación: Una vez terminada tanto la implementación como el documento se realiza una presentación con un resumen del proyecto.
- Tiempo invertido: 8 horas.
- Participantes: Soledad Martín Morales.

El total de horas estimadas para el diseño es: 68 horas.

La duración real del proyecto es de **339h**.

Como he comentado anteriormente tanto la planificación inicial como la planificación real está definida en horas en vez de días ya que no son 8 exactamente las horas invertidas en el trabajo, si no que esta cifra fluctúa dependiendo del día por motivos laborales. Por este motivo veo más efectivo realizar la estimación en horas.

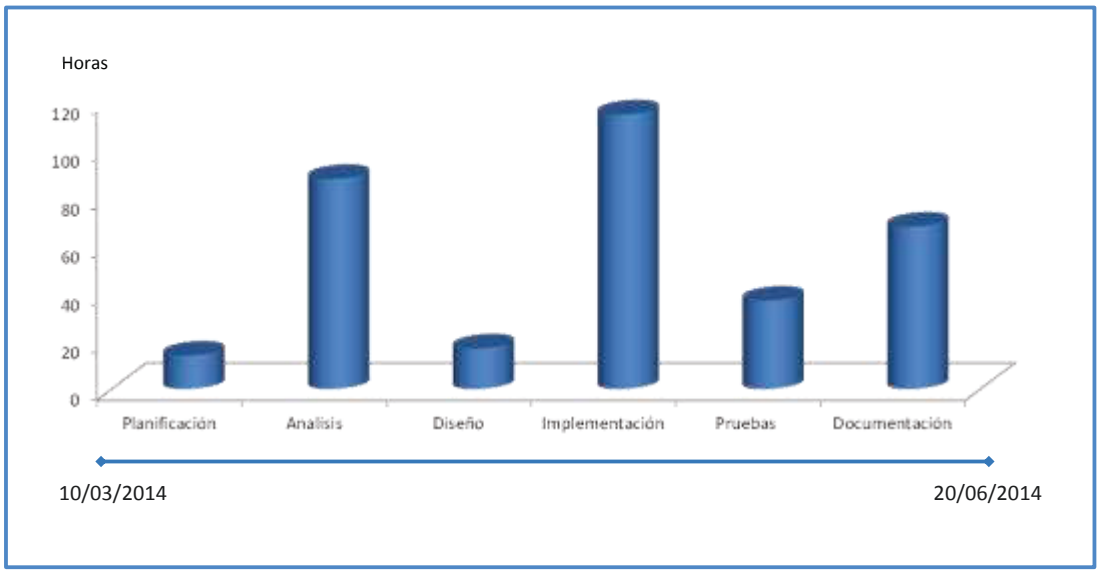
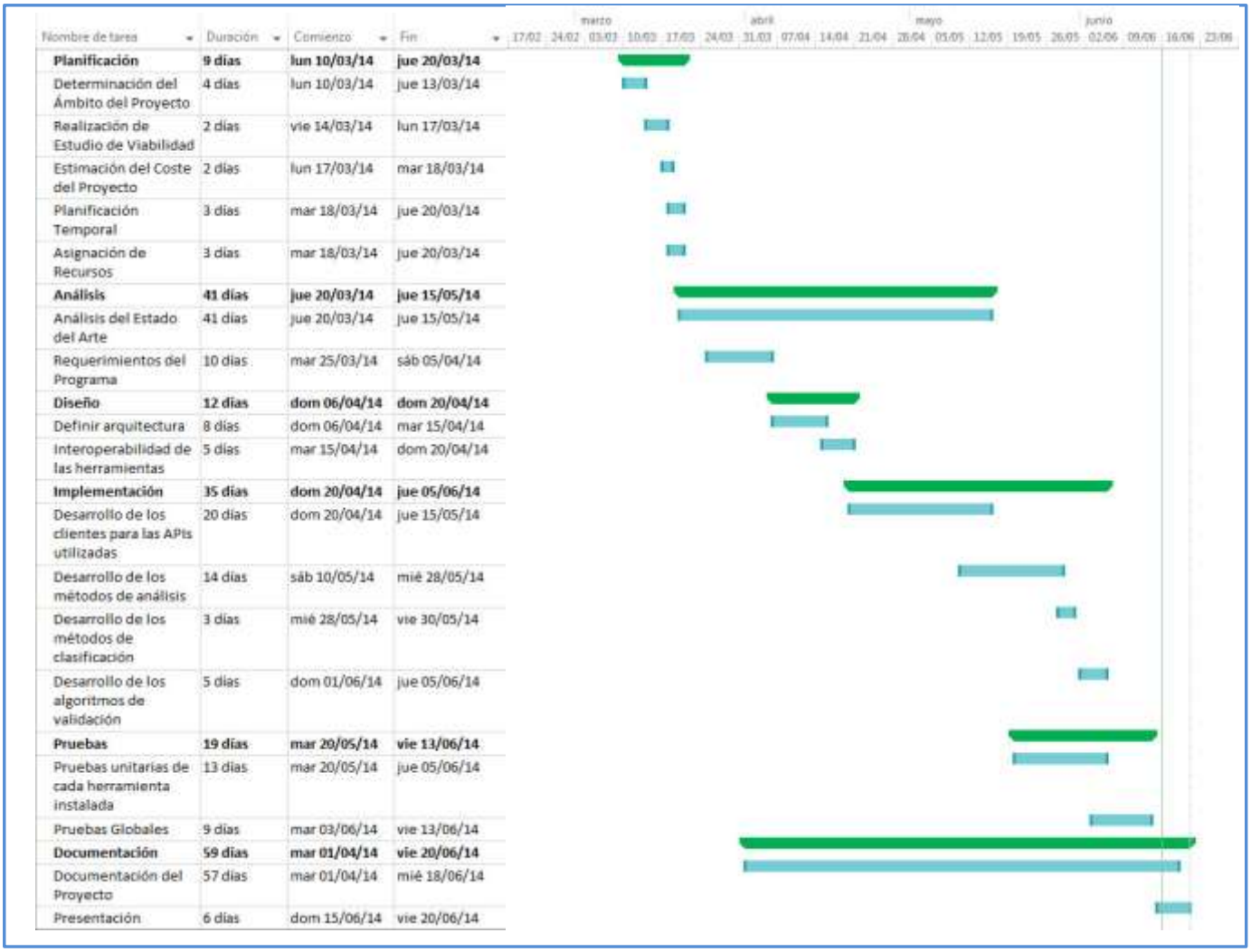


Ilustración 38. Ejecución de Horas en cada Fase del Proyecto



Nombre de tarea	Duración	Comienzo	Fin
Planificación	**9 días**	**lun 10/03/14**	**jue 20/03/14**
Determinación del Ámbito del Proyecto	4 días	lun 10/03/14	jue 13/03/14
Realización de Estudio de Viabilidad	2 días	vie 14/03/14	lun 17/03/14
Estimación del Coste del Proyecto	2 días	lun 17/03/14	mar 18/03/14
Planificación Temporal	3 días	mar 18/03/14	jue 20/03/14
Asignación de Recursos	3 días	mar 18/03/14	jue 20/03/14
Análisis	**41 días**	**jue 20/03/14**	**jue 15/05/14**
Análisis del Estado del Arte	41 días	jue 20/03/14	jue 15/05/14
Requerimientos del Programa	10 días	mar 25/03/14	sáb 05/04/14
Diseño	**12 días**	**dom 06/04/14**	**dom 20/04/14**
Definir arquitectura	8 días	dom 06/04/14	mar 15/04/14
Interoperabilidad de las herramientas	5 días	mar 15/04/14	dom 20/04/14
Implementación	**35 días**	**dom 20/04/14**	**jue 05/06/14**
Desarrollo de los clientes para las APIs utilizadas	20 días	dom 20/04/14	jue 15/05/14
Desarrollo de los métodos de análisis	14 días	sáb 10/05/14	mié 28/05/14
Desarrollo de los métodos de clasificación	3 días	mié 28/05/14	vie 30/05/14
Desarrollo de los algoritmos de validación	5 días	dom 01/06/14	jue 05/06/14
Pruebas	**19 días**	**mar 20/05/14**	**vie 13/06/14**
Pruebas unitarias de cada herramienta instalada	13 días	mar 20/05/14	jue 05/06/14
Pruebas Globales	9 días	mar 03/06/14	vie 13/06/14
Documentación	**59 días**	**mar 01/04/14**	**vie 20/06/14**
Documentación del Proyecto	57 días	mar 01/04/14	mié 18/06/14
Presentación	6 días	dom 15/06/14	vie 20/06/14

Ilustración 39. Diagrama de Gantt, Ejecución Final

9 PRESUPUESTO

En el presupuesto se tienen en cuenta tanto el número de horas invertidas por las personas que realizan el proyecto como las herramientas utilizadas para el funcionamiento de este.

Tal y como se ha visto en el apartado anterior en la fase de planificación se crea un presupuesto inicial y una vez finalizado el proyecto se hace una valoración económica real.

Para valorar el coste que supone cada una de las personas que realizan el proyecto, además de tener en cuenta su tarifa, es necesario prever los costes de la seguridad social, las retenciones, etc.

La persona que realiza el proyecto tiene que ser un Ingeniero Informático Junior con conocimiento en programación Java. Además tiene que tener la capacidad de gestionar el proyecto ya que va a estar inmerso en el desde la planificación hasta la documentación. No es necesario el conocimiento de las herramientas que se van a utilizar aunque si es recomendable, pero ese conocimiento se irá adquiriendo a lo largo del proyecto.

La persona encargada de tutelar el proyecto tiene que ser Ingeniero Informático Senior, profesor de la Universidad Carlos III y tener experiencia en la gestión de los proyectos de fin de grado. Además tiene que tener un gran conocimiento del *Data Mining* y de todo el tema relacionado con el análisis de textos.

Se adjunta como (Anexo I. Calculo Salario Profesionales) el cálculo del salario neto de los profesionales asignados al proyecto para comprobar su viabilidad y el cálculo del coste/hora de cada uno.

Teniendo en cuenta la información recogida en el anexo el coste/hora de cada profesional es el siguiente:

- Ingeniero Jr. : 10,23€
- Ingeniero Sr. : 22,73€

Ambos presupuestos están estimados a partir de las horas de trabajo realizadas. Se elige realizar la estimación en horas en vez de días porque por motivos laborales no se puede invertir 8 horas diarias en el proyecto.

El coste material vendrá definido por el coste de las herramientas tanto hardware como software necesarias para realizar el proyecto.

Los costes relacionados con las dietas, los traslados, costes de estructura, etc. serán asumidos por las personas que realizan el proyecto por lo que no supondrán ningún coste adicional.

9.1 PRESUPUESTO INICIAL

9.1.1 Coste Personal

Para calcular el coste del personal se han tenido en cuenta las dos personas relacionadas con el proyecto, la función de cada una de ellas y el número de horas invertidas.

La valoración económica inicial para estos dos perfiles, teniendo en cuenta el número de horas invertidas es la siguiente:

	Perfil	Tarifa/Hora	Horas Totales	Coste Total
Recurso 1	Ingeniero Informático Jr.	10,23 €	300	3.069 €
Recurso 2	Ingeniero Informático Sr.	22,73 €	24	546 €
			Coste Total Personal	**3.615 €**

Ilustración 40. Estimación Inicial, Coste Personal

9.1.2 Coste Herramientas

Al coste del personal hay que añadirle el coste de todas las herramientas necesarias para el proyecto. Todos estos costes no incluyen IVA.

En la tabla se mostrará el producto, su precio, el periodo de amortización en meses y el tiempo que ha sido utilizado. En el caso de las herramientas que tienen un coste variable, si lo utilizas para fines didácticos y de investigación te lo ofrecen de manera gratuita.

En este presupuesto se asocia el coste de la herramienta al uso que se hace de ella.

	Precio	Periodo de Amortización (meses)	Tiempo de Uso (meses)	Coste Total
Ordenador Portatil ASUS A53E	545 €	36 €	3	45 €
Windows 7 Profesional	105 €	36 €	3	9 €
Microsoft Office	539 €	36 €	3	45 €
Freeling	0 €	-	2	0 €
Normalizador RAE	0 €	-	2	0 €
Wordnet	0 €	-	2	0 €
			Coste Total Herramientas	**99 €**

Ilustración 41. Estimación Inicial, Coste Herramienta

9.1.3 Coste Total

El coste total del proyecto es la suma del coste del personal más el coste de la herramienta. A esta cantidad también hay que añadirle un porcentaje de beneficio y el resultado se mostrará con IVA y sin IVA. El IVA en la actualidad supone un 21%.

Concepto	Coste
Personal	3.615 €
Herramientas	99 €
Beneficio	20%
sin IVA	4.456 €
con IVA (21%)	**5.392 €**

Ilustración 42. Presupuesto Inicial, Coste Total

Por lo tanto el proyecto, según el presupuesto inicial tendría un coste de **5.392€.**

9.2 COSTE REAL

Se produce una variación entre la planificación inicial y la real de 39 días de más, y esto repercute en el presupuesto real.

Además se han utilizado otras herramientas distintas a las iniciales pero esto no ha tenido repercusión.

9.2.1 Coste Personal

En el presupuesto real se toman las mismas premisas que en el presupuesto inicial en cuanto a perfiles, valoración económica en horas, etc.

La valoración económica para estos dos perfiles, teniendo en cuenta el número de horas invertidas en la realidad es la siguiente:

	Perfil	Tarifa/Hora	Horas Totales	Coste Total
Recurso 1	Ingeniero Informático Jr.	10,23 €	339	3.468 €
Recurso 2	Ingeniero Informático Sr.	22,73 €	19	432 €
			Coste Total Personal	**3.900 €**

Ilustración 43. Coste Real, Coste Personal

9.2.2 Coste Herramientas

Tal y como se ha comentado anteriormente las aplicaciones utilizadas finalmente han variado en relación con las iniciales pero esto no ha supuesto ningún cambio económico en este apartado.

	Precio	Periodo de Amortización	Tiempo de Uso (meses)	Coste Total
Ordenador Portatil ASUS A53E	545 €	36 €	4	61 €
Windows 7 Profesional	105 €	36 €	4	12 €
Microsoft Office	539 €	36 €	4	60 €
Apicultur	0 €	-	2	0 €
Normalizador RAE	0 €	-	2	0 €
API Twitter	0 €	-	2	0 €
			Coste Total Herramientas	**132 €**

Ilustración 44. Coste Real, Coste Herramienta

9.2.3 Coste Total

Igual que pasaba en el presupuesto inicial el coste total real del proyecto es la suma del coste del personal más el coste de la herramienta. A esta cantidad también hay que añadirle un porcentaje de beneficio y el resultado se mostrará con IVA y sin IVA. El IVA en la actualidad supone un 21%.

Concepto	Coste
Personal	3.900 €
Herramientas	132 €
Beneficio	15%
sin IVA	4.838 €
con IVA (21%)	**5.610 €**

Ilustración 45. Total Coste Real

Con la finalidad de que no tengan una gran repercusión económica los problemas que se han encontrado a lo largo del desarrollo del proyecto, se ha disminuido el margen de beneficio. Pasa de ser de un 20% a un 15%. Se asumen los costes generados por los errores ya que el proyecto se va a entregar en una fecha posterior a la estimada.

Por lo tanto, el coste final del proyecto es de **5.610€**

A continuación se muestra una tabla comparativa del presupuesto inicial y del presupuesto final:

Concepto	Coste Inicial	Coste Final
Personal	3.615 €	3.900 €
Herramientas	99 €	132 €
Beneficio	20%	15%
sin IVA	4.456 €	4.838 €
con IVA (21%)	**5.392 €**	**5.610 €**

Diferencia	**218 €**

Ilustración 46. Comparativa Costes

La diferencia entre la estimación inicial y el coste real es de 218€. A pesar de que el coste sea mayor el beneficio será menor ya que hemos tenido que disminuir su porcentaje. Aun así se saca un beneficio del 15% en el proyecto.

10 PROTECCIÓN DE DATOS

Una preocupación muy extendida en la actualidad es la protección de datos en las redes sociales, es conveniente comentar que para este proyecto se ha cumplido con Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal cuyo objeto es el siguiente *"La presente Ley Orgánica tiene por objeto garantizar y proteger, en lo que concierne al tratamiento de los datos personales, las libertades públicas y los derechos fundamentales de las personas físicas, y especialmente de su honor e intimidad personal y familiar"*

Todos los datos obtenidos a partir de la herramienta de Twitter han sido utilizados única y exclusivamente para analizar el sentimiento de su contenido. Además se ha partido de datos públicos (solo se obtenían los tweets y el usuario que lo había redactado).

11 CONCLUSIONES

En este proyecto se ha descrito la creación de un motor en Java que muestra el análisis de sentimiento de los Tweets obtenidos a través de Twitter. Para ello se han utilizado una serie de herramientas externas y se han creado métodos propios para el tratamiento de los datos.

El análisis de sentimientos es una tarea muy importante hoy en día en toda red social y es muy útil, sobre todo, para conocer la opinión de la gente acerca de un tema concreto.

El Tweet analizado pasa por una serie de módulos en los que se han ido tratando cada una de las palabras que lo compone. Se ha normalizado el Tweet, se han eliminado palabras nulas y se han realizado otra serie de pasos hasta que, finalmente, se ha valorado numéricamente cada una de las palabras y se ha generado un algoritmo que ofrecía un número como solución. Este valor determinaba el sentimiento del Tweet que podía ser positivo, negativo o neutro.

Este proyecto ha supuesto un gran reto para mí y me ha servido tanto para aprender mucho acerca del Data Mining y el análisis de la información como para aprender a desarrollar un proyecto desde el inicio.

He tenido que realizar un trabajo muy extenso de documentación y de conceptualización ya que es un tema en auge actualmente, con mucha información, pero con un mundo entero por descubrir. Existe mucha documentación en inglés, debido a su sencillez, pero en España aún queda mucho trabajo por hacer.

Con el desarrollo del proyecto también he adquirido muchos conocimientos acerca del ciclo de vida un proyecto. Todas las fases previas a la implementación que son tan necesarias para que el trabajo sea productivo y la labor de documentación, tan necesaria también, para reflejar todo el trabajo realizado. Otro apartado importante y que me ha aportado gran conocimiento ha sido la planificación y el presupuesto del proyecto.

Todos los conocimientos que he ido adquiriendo durante el proyecto y todas las horas invertidas buscando información han ido generando en mí una inquietud por seguir trabajando en esta línea, tanto en temas relacionados con la gestión de los datos en la red, como en la gestión de proyectos.

Espero que en un futuro, esta haya sido la base de un gran conocimiento.

12 TRABAJOS FUTUROS

A medida que se ha ido desarrollando el proyecto se han encontrado varios puntos de análisis que no ha sido posible implementar en esta entrega pero que se proponen para realizar en el futuro con la finalidad de aportar valor adicional a éste. Las mejoras propuestas son:

- En la primera fase del tratamiento del tweet (Ver apartado 4.1. Tratamiento de palabas especiales) se sacaba el *#Hashtag* del tweet y luego se le daba una valoración neutra. La primera mejora que se propone es el tratamiento de esta palabra ya que se considera un punto clave dentro del análisis de cada tweet.
- En la segunda fase (Ver apartado 4.2. Eliminación de caracteres) se procedía a la eliminación de la palabra completa en el caso de que esta fuese una URL. Se propone como mejora la gestión de las URLs, de manera que se pueda comprobar a que página te lleva y poder valorar la información.
- Aumentar el número de herramientas en el análisis de los datos para mejorar la precisión. Las herramientas que se proponen inicialmente son:
  - **Freeling.** Tal y como se ha comentado en el estado del arte (Ver apartado 2.4. Herramientas de apoyo para el análisis de datos), Freeling es una librería de código abierto para el procesamiento multilingüe automático. Ofrece a los desarrolladores funciones de análisis y anotación lingüística de textos. Para el uso de esta herramienta se recomienda la realización del proyecto en Linux.
  - **Wordnet.** Es una base de datos léxico-conceptual estructurada en forma de red semántica, es decir, compuesta de unidades léxicas y relaciones entre ellas. La versión 3.0 está en Español y también hay muchas más facilidades si se ejecuta sobre Linux.
- En el mismo apartado se trataba la eliminación de las palabras nulas. En este caso se ha hecho de forma manual con una lista de palabras pero se propone como mejora pasar todas estas palabras por una herramienta que categorice la palabra y en el caso de ser una categoría que asegura que la palabra no va a portar valor (una preposición, por ejemplo), esta podría ser eliminada. El problema que surge al realizar la modificación es el "gasto" que se hace de la herramienta, ya que cada herramienta tiene un límite de tokens que se pueden analizar en un tiempo determinado, sería necesario valorar su viabilidad.
- Clasificar el resultado en cinco categorías en vez de en tres como hay ahora para hacer un análisis más exhaustivo del sentimiento. Estas categorías serían muy positivo, positivo, neutro, negativo y muy negativo. Para crear esta modificación habría que estudiar el límite (en número) entre positivo y muy positivo, y negativo y muy negativo.

Como he comentado en el apartado anterior es un tema en auge actualmente por la cantidad de mejoras que se podrían realizar es infinita. Las aquí descritas son una pequeña muestra para iniciar la mejora del proyecto.

13 GLOSARIO DE ACRÓNIMOS

API → Application Programming Interface)

BBVA → Banco Bilbao Vizcaya Argentaria

CSV → comma-separated values

HTTP → Hypertext Transfer Protocol

IOS → iPhone OS

IVA → Impuesto de Valor Añadido

PLN → Procesadores de Lenguaje Natural

RAE → Real Academia Española

SaaS -> Software as a solution

SDK → Software Development Kit

SEPLN → Sociedad Española para el Procesamiento del Lenguaje Natural

SLA -> Acuerdo a nivel de Servicio

TALP -> Centre de Tecnologies i Aplicacions del Llenguatge i la Parla

TFG→ Trabajo de Fin de Grado

UPC -> Universitat Politècnica de Catalunya

URL → Uniform Resource Locator

XML → eXtensible Markup Language

14 BIBLIOGRAFÍA

A. Montejo-Ráez, E. M.-C.-V.-L. (s.f.). *Detección de la polaridad en citas periodísticas: una solución no supervisada.* Obtenido de http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/viewFile/4551/2717

Alegría, I., Etxeberría, I., & Labaka, G. (2013). Una cascada de transductores simples para normalizar tweets.

Alonso i Alemany, A. (2005). *Herramientas Libres para Procesamiento de Lenguaje Natural*. Obtenido de http://www.cs.famaf.unc.edu.ar/~laura/freeNLP

Álvarez, M. Á. (s.f.). *Introducción a la API de Twitter.* Obtenido de http://www.desarrolloweb.com/articulos/intro-api-twitter-curl.html

Baraldi, L. (05 de 2012). *leobaraldi.com*. Recuperado el 04 de 2014, de http://www.leobaraldi.com.ar/2012/05/usar-la-api-de-facebook-por-donde-comenzar/

BBVA. (s.f.). *ActiBVA*. Obtenido de http://www.actibva.com/

Bernardos, M. d. (2003). *Marco metodológico para la construcción de sistemas de generación de lenguaje natural.* Obtenido de http://oa.upm.es/183/1/10200308.pdf

Berzal, F. (s.f.). *Ciclo de Vida de un Sistema de Información.*

Berzal, F. (s.f.). *El ciclo de vida de un sistema de información*. Obtenido de http://elvex.ugr.es/idbis/db/docs/lifecycle.pdf

Botía, J. A. (2009). *Preprocesado de Datos.* Obtenido de Universidad de Murcia: file:///C:/Users/Solete/Downloads/tiia0809_slides_prep.pdf

Cardenas Montes, M. (2013). *Preprocesado de Datos.* Obtenido de Centro de Investigaciones Energéticas Medioambientales y Tecnológicas,: http://wwwae.ciemat.es/~cardenas/curso_MD/preprocesado_datos.pdf

Castells, P. (2012). *La Web Semántica.* Obtenido de http://arantxa.ii.uam.es/~castells/publications/castells-uclm03.pdf

Computerworld. (16 de Junio de Junio 2014). *IBM Watson y Genesys impulsan juntas la relación empresa-cliente.* Obtenido de http://www.computerworld.es/sociedad-de-la-informacion/ibm-watson-y-genesys-impulsan-juntas-la-relacion-empresascliente

Daedalus. (2013). *Sentimentalytics Home Page.* Obtenido de https://sentimentalytics.com/inicio#.U1uKpPl_ulU

Daedalus. (s.f.). *Textalytics*. Obtenido de Home Page: http://textalytics.com/inicio

De Pablo, C. M. (Octubre 2013). *La manera más sencilla de incorporar procesamiento semántico a tus aplicaciones.* Obtenido de

http://www.slideshare.net/Daedalus_SA/webinar-textalytics-meaning-as-a-service-daedalus-8-octubre-2013

Díaz Esteban, A., Alegria Loinaz, I., & Villena Román, J. (2013). XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural.

Fernández Montraveta, A. (s.f.). *La construcción del WordNet 3.0 en Español.* Obtenido de http://www.academia.edu/972301/La_construccion_del_WordNet_3.0_en_espanol

INTECO. (2009). *Ingeniería del Software- Metodologías y ciclo de vida.* Obtenido de https://www.inteco.es/file/N85W1ZWFHifRgUc_oY8_Xg

Landazabal, A. (2008). *Ingenieriía de los requisitos*. Obtenido de http://www.educaplay.com/es/recursoseducativos/1055240/html5/ingenieria_de_los_requisitos.htm#!

León Guzmán, E. (2010). *Preprocesamiento de Datos.* Obtenido de http://disi.unal.edu.co/profesores/eleonguz/cursos/md/presentaciones/Sesion3_preprocesamiento.pdf

Marín Diazaraque, J. M. (2009). *Introducción al Data Mining.* Obtenido de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/introduccion-DM.pdf

Marrero, M. (2013). *Evaluación en Recuperación de la Información.*

Moler, C. (s.f.). *Matlab*. Obtenido de http://www.mathworks.es/products/matlab/

Mosquera, A., & Moreda, P. (2013). DLSI en Tweet-norm 2013: Normalizacion y Tweets en Español.

Padró, L. (s.f.). *Analizadores Multilingües en Freeling.* Recuperado el 20 de Enero de 2014, de Linguamatica: http://nlp.lsi.upc.edu/papers/padro11.pdf

Padró, L. (Octubre 2013.). *FreeLing User Manual.* Obtenido de http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf

Page., W. H. (s.f.). Obtenido de http://wordnet.princeton.edu

Pascual, I. V. (2012). *La Web Semantica.* Obtenido de http://lawebsemantica.weebly.com/

ProfilerPlus. (s.f.). *Home Page Social Science Autimation.* Obtenido de http://socialscience.net/tech/ProfilerPlus.aspx

Pyle. (1999). *Preprocesar los datos.*

Rodríguez Gil-Ortega, M. J. (s.f.). *Wrappers.* Obtenido de Universidad Politécnica de Madrid.: http://sinbad.dit.upm.es/docencia/doctorado/curso0304/Wrappers.pdf

Sánchez Suarez, J. M. (s.f.). *Análisis de los sentimientos en twitter con el soporte de Apicultur. .* Obtenido de Apicultur:

http://www.adictosaltrabajo.com/tutoriales/tutoriales.php?pagina=analisisSentimientosTwitterApicultur

SEPLN. (2013). Obtenido de http://nil.fdi.ucm.es/sepln2013/

SEPLN. (s.f.). *Home Page.* Obtenido de http://www.sepln.org/

SEPLN. (s.f.). *TASS - Workshop on Sentiment Analysis at SEPLN.* Obtenido de http://rua.ua.es/dspace/bitstream/10045/27862/1/PLN_50_04.pdf

Sinnexus. (s.f.). Obtenido de Business Intelillence, Informática Estratégica: http://www.sinnexus.com/empresa/index.aspx

Troyano, J. A. (s.f.). *Definición de Wordnet. .* Obtenido de http://www.lsi.us.es/~troyano/documentos/wordnet.pdf

Vallejo, C. (2013). *Aprendizaje por Proyectos y TIC.* Obtenido de http://recursostic.educacion.es/observatorio/web/es/component/content/article/1057-aprendizaje-por-proyectos-y-tic?start=3

Villar Rodriguez, E., Garcia Serrano, A., & González Rodríguez, M. (2013). Análisis lingüístico de expresiones negativas en tweets en español.

Waikato, M. L. (s.f.). *WEKA*. Obtenido de http://www.cs.waikato.ac.nz/ml/weka/

Witten, E. F. (2005). Data Mining: Practical Machine Learning Tools and Techniques, Second Edition.

ANEXO I. CALCULO SALARIO PROFESIONALES

A través de ActiBVA (BBVA), que es una comunidad impulsada por BBVA y desarrollada por Weblogs SL cuyo objeto es acercar la información financiera a todos los usuarios se ha realizado el cálculo del salario neto anual de cada profesional asignado para comprobar su viabilidad en el proyecto.

Los cálculos son los siguientes:

Ingeniero Informático Jr.

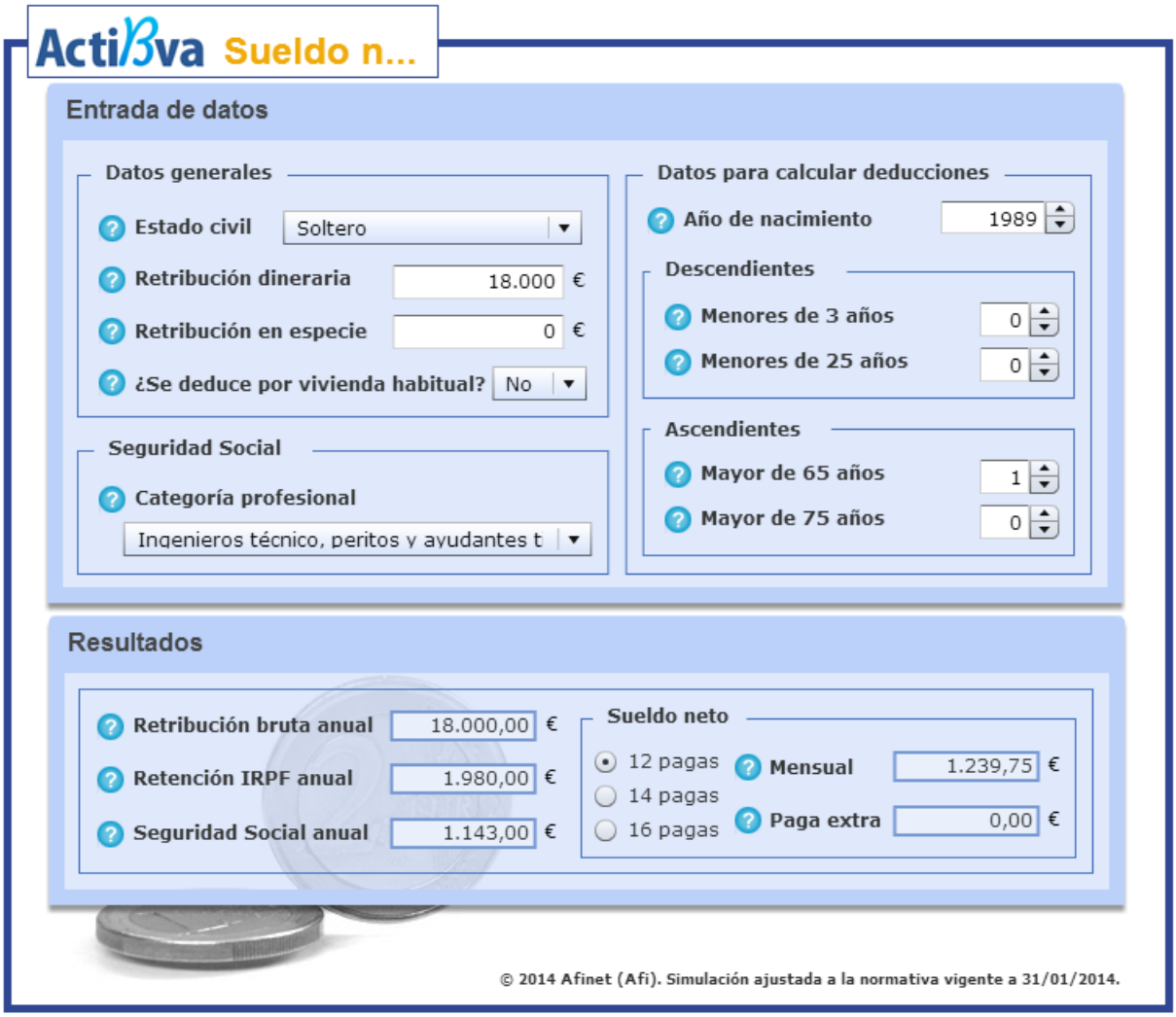


Ilustración 47. Calculo Salario Informático Jr.

Ingeniero Informático Sr.

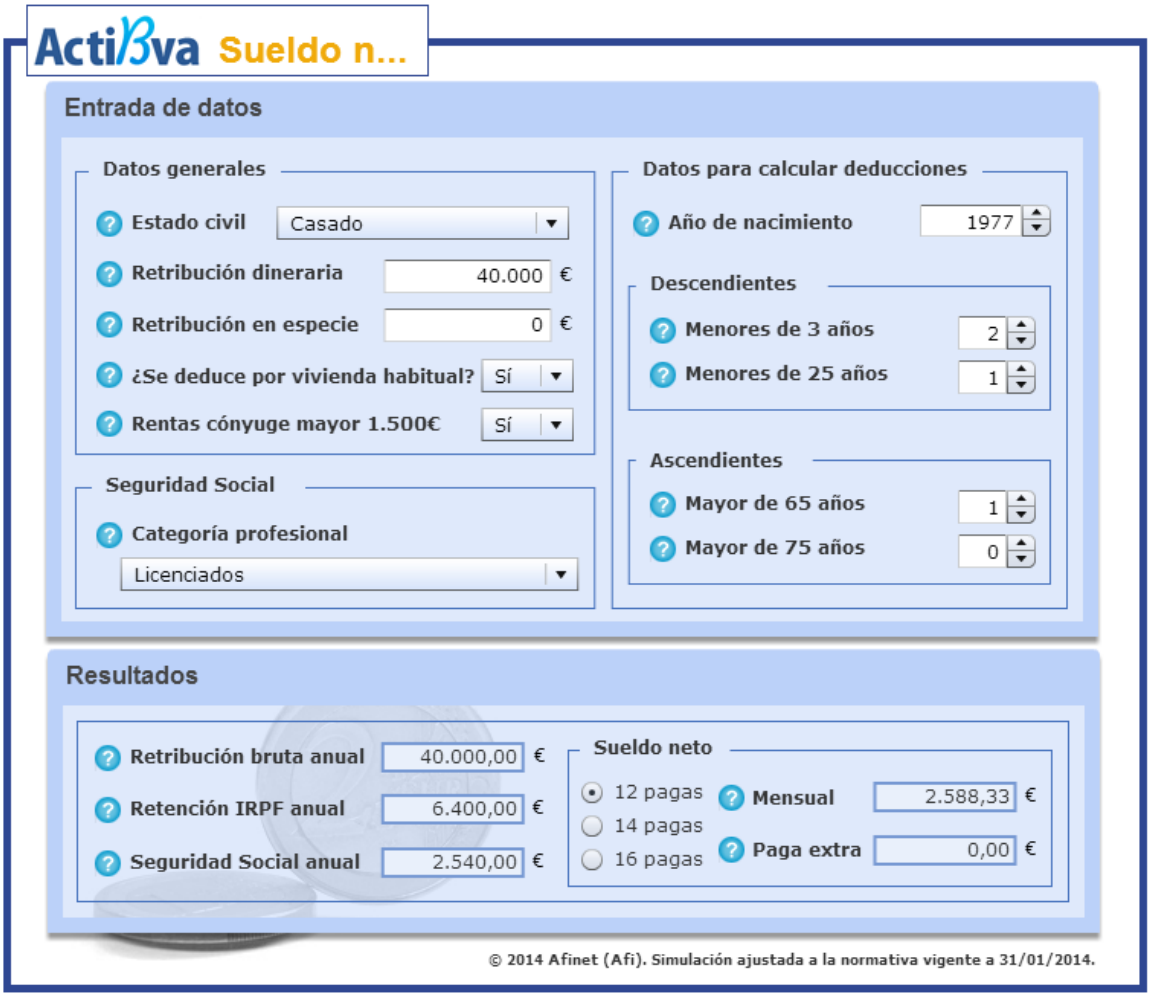


Ilustración 48. Cálculo Salario Informático Sr.

El sueldo que nos muestra este cálculo es mensual, el resultado que se ha obtenido parece óptimo por lo que a continuación se va a estimar en horas.

Para este cálculo se utiliza un estándar de horas anuales estipulado en 1760. Este número de horas se debe a que un año tiene 365, de los que se trabajan 220 aproximadamente (entre fines de semana, festivos y vacaciones) y el número de horas de trabajo son 8, por lo tanto:

(365 – 145 (festivos y vacaciones)) * 8 = 1760.

Teniendo en cuenta este estándar, la tarifa es la siguiente:

- **Ingeniero Jr. : 18.000/1760 = 10,23€**
- **Ingeniero Sr. : 40.000/1760 = 22,73€**